The Predictive Power of Social Media Data

Buus Lassen, Niels

Document Version
Final published version

DOI:
10.22439/phd.29.2023

Publication date:
2023

License
Unspecified

Citation for published version (APA):
Buus Lassen, N. (2023). *The Predictive Power of Social Media Data*. Copenhagen Business School [Phd]. PhD Series No. 29.2023 https://doi.org/10.22439/phd.29.2023

COPENHAGEN BUSINESS SCHOOL
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DANMARK

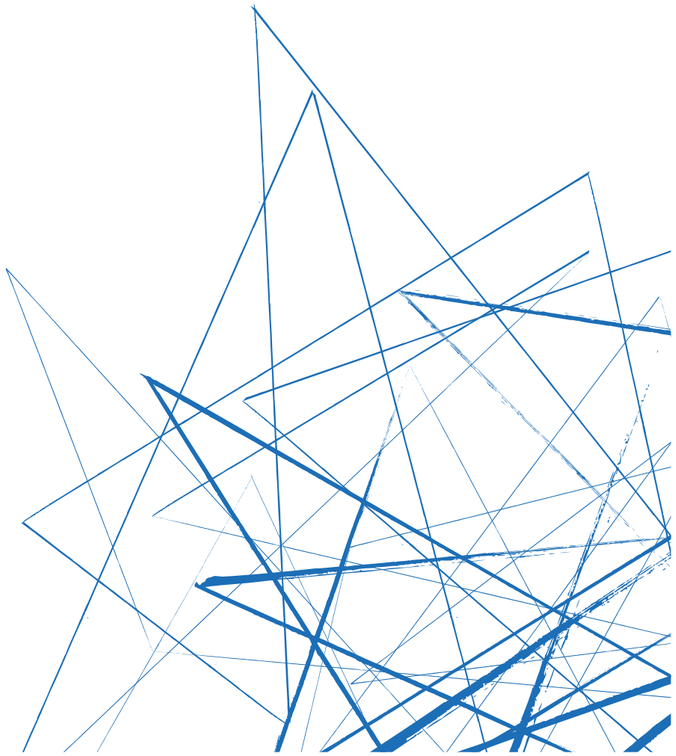
WWW.CBS.DK

ISSN 0906-6934

Print ISBN: 978-87-7568-199-0
Online ISBN: 978-87-7568-200-3

DOI: https://doi.org/10.22439/phd.29.2023

THE PREDICTIVE POWER OF SOCIAL MEDIA DATA



PhD Series 29.2023

Niels Buus Lassen

THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

DIGI

PhD Series 29.2023

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

Niels Buus Lassen
Primary supervisor: Ravi Vatrapu
Secondary supervisor: Lisbeth la Cour

BM PhD School
Department of Digitalization
Copenhagen Business School

Niels Buus Lassen
THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

First edition 2023
Ph.D. Series 29.2023

ISSN 0906-6934

Print ISBN: 978-87-7568-199-0
Online ISBN: 978-87-7568-200-3

DOI: https://doi.org/10.22439/phd.29.2023

Acknowledgements

I would like to thank Industriens Fond for financing my PhD under their 6,2 million DKK donation for big social data research at Copenhagen Business School. I would also like to thank my two supervisors, Prof. Ravi Vatrapu and Prof. Lisbeth la Cour, for invaluable help and support during my PhD journey. A special thanks to Prof. Jacob Nørbjerg, who was invaluable in helping me form my Kappa in the last phase of my PhD journey. Many thanks to Prof. Erik Wong, who also helped with the direction of my Kappa. Many thanks to Dr. Laura Pascal, Dr. Benjamin Flesch, Evnita Karlsson and Dr. Mimi Hou for invaluable help with proofreading. My time as visiting scholar at Nuffield College and Economics Dept. at Oxford University was invaluable for my big social data research, and I especially thank Prof. Sir David Hendry, Prof. Bent Nielsen and Prof. Anders Kock.

English Abstract

Predictive analytics using social media data is a relatively new field, being widely recognized at first due to "Predicting the Future with Social Media" of Asur and Huberman (2010). Their article proposed a model that could predict the revenues of Hollywood movies using Twitter data as input. This model has proven more accurate than the golden standard for movie revenue prediction – the Hollywood Stock Exchange – being thus a betting market for Hollywood movie performances and considered a good prediction market example.

Since 2010, predictive analytics using social media data have developed extensively. In line with this trend, this PhD thesis presents pioneering research on the first Twitter-based prediction sales model, which explains why Twitter data can predict iPhone sales. It also demonstrates that similar predictive sales models can be built using Facebook and Google search data. The methodology is based on customer journey models, which are the main conceptual models for categorizing all social media and web search data in a sales context. Based on this methodology, an investor journey model for categorizing web search data in a financial market context is proposed.

This thesis is structured as a collection of five research papers, written sequentially. The papers shows five predictive models. A successful predictive model for iPhone sales using Twitter data, and a successful predictive model for H&M using Facebook data. The predictive sales model for Mikkeller beers using Facebook and Google data is not successful. Social data are then conceptualized to build a general model for predictions based on social media data. The first four predictive models use the marketing-based customer journey models to explain the association mechanism linking input data to predicted sales. This is further developed to create an investor journey model, which is then used to build a

predictive model for investor behaviour using Google search data as input.

In terms of the results, it is shown, that it is possible to predict iPhone sales with Twitter data and H&M sales with Facebook data. Conversely, it is also shown, that it is not possible to predict Mikkeller beer sales with either Google or Facebook data. Research into the reasons for these predictive modelling successes and failures provides important insights about both the nature and size of social data, which are determining their potential use in predictive models. These insights ultimately led to the development of Figure 10: Social Filtering Model. This model shows how filtering is determining the content we consider and measure on different social platforms. This model is the most important contribution of this PhD thesis, as it lays the foundation for a more scientific discussion on the potential uses for social data. The Social Filtering Model also led to the development of the chapter 8. Predictive Modelling Framework.

The fifth model shows the predictive power of Google searches for Apple stock volatility applied to Apple stock investor behaviour, based on the customer journey concepts developed in the first four predictive models. While the model failed to model the Apple stock price, visual graphs of Google searches for the Amazon stock symbol showed a higher correlation with the Amazon stock prices than Apple stocks. Therefore, the method in the fifth model can probably be used for modelling stock prices for stocks other than Apple. The main contribution of the fifth model is the investor journey model, explaining why Google search data have predictive power for investor behaviour. Another contribution is the identification of private and professional investors' different uses of Google searches and the high importance of stock symbols, among all the stock related Google searches for predictive modelling of investor behaviour.

This PhD thesis contributes to the academic discussion on identifying the

predictive power in social data for the predictive modelling of sales and investor behaviour. The key contributions are proposing a new conceptual Social Filtering Model for social data, Figure 10: Social Filtering Model., explaining how human behaviour on social media and search engines are filtered differently, and how this determines the potential use for data in a practical modelling Framework, in chapter 8. Predictive Modelling Framework.

Dansk Abstrakt

Prædiktiv analyse på basis af sociale medie data, er et relativt nyt forskningsområde. Det blev først bredt anerkendt, da artiklen "Predicting the Future with Social Media" blev skrevet af Asur og Huberman (2010). Deres artikel foreslog en model, som kunne prædiktere omsætningen for en del Hollywood film, med Twitter data som input. Deres model viste sig, at være mere nøjagtig end den gyldne standard for prædiktering af film omsætning – the Hollywood Stock Exchange – som er et marked for væddemål, meget lig Oddset på Danskespil.dk. På Hollywood Stock Exchange laver folk væddemål med odds, hvor de gætter på omsætningen for en del Hollywood film. Den kollektive visdom i alle de væddemål, har dannet ramme for prædiktering af Hollywood film omsætning i mange år, og bliver anset for at være et godt eksempel på prædiktive markeds data.

Siden 2010, har prædiktiv analyse med sociale medie data udviklet sig meget omfattende. I tråd med denne udvikling, præsenterer denne PhD afhandling banebrydende forskning med den første Twitter baserede salgs prædikterings model, som forklarer hvorfor Twitter data kan prædiktere salg, i form af iPhone omsætning. Denne PhD afhandling demonstrerer også, at lignende prædiktive salgs modeller kan blive bygget med Facebook og Google søgnings data.

Metoden I disse modeller er baseret på kunderejse modeler, som er de centrale konceptuelle modeller til at kategorisere alle sociale medie og web søgnings data i en salgs kontekst. Baseret på denne metodologi bliver der også præsenteret en investor rejse model til at kategorisere web søgnings data i en finans marked kontekst for Apple aktien.

Denne PhD afhandling er struktureret som en samling af fem forsknings artikler, der er skrevet sekventielt og viser udviklingen af fem prediktive modeller. En succesfuld prædiktiv model for iPhone omsætningen, på basis af Twitter data, og en succesfuld prædiktiv model for H&M omsætningen på basis af Facebook data. Den prædiktive salgs model for Mikkeller øl med brug af Facebook og Google søgnings data er ikke succesfuld. Derefter konceptualiseres sociale data, og der bygges en generel prædiktiv model baseret på sociale medie data. De første fire artikler bruger en marketing baseret kunderejse model til at forklare associationerne, der forbinder de sociale data med det prædiktive salg. Dette bliver videre udviklet i den femte model, hvor der skabes en investor rejse model, som bliver brugt til at bygge en prædiktiv model for investor adfærd med Google søgnings data som input.

Resultaterne viser, at det er muligt at prædiktere iPhone omsætningen med Twitter data, og H&M omsætningen med Facebook data. Det vises også, at det ikke er muligt at prædiktere Mikkeller øl salg med hverken Google søgnings eller Facebook data. Forskning i grundene bag disse succeser og fiaskoer har givet nogle vigtige indsigter i både arten og størrelsen af sociale data, som bestemmer den potentielle brug i prædiktive modeller. Disse indsigter ledte frem til udviklingen af den sociale filtrerings model, som præsenteres i kapitel 7, Figure 10: Social Filtering Model. Denne model viser hvordan filtrering bestemmer det indhold vi kan se og måle på, på de forskellige sociale platforme. Denne model er det vigtigste bidrag for denne PhD, da den ligger fundamentet for en mere

videnskabelig diskussion af den potentielle brug af sociale data. Den sociale filtrerings model ledte også frem til en prædiktiv model bygge anvisning, som også præsenteres i kapitel 8, Table 14, Predictive Modelling Framework.

Den femte model viser den prædiktive kraft i Google søgninger for Apple aktiens volatilitet. Denne forskning er baseret på Apple investorers adfærd på Google søgninger, og kunderejse koncepterne fra de første fire modeller udviklet videre som en investor rejse model. Selv om modellen ikke kunne modellere Apple aktiepris, viste visuelle grafer for Google søgninger af Amazon aktiesymbolet, en højere korrelation med Amazon aktieprisen, end for Apple aktien. Metoden i den femte artikel kan derfor godt bruges til modellering af aktiepriser i nogle tilfælde. Hoved bidraget i den femte artikel er investor rejse modellen, der forklarer hvorfor Google søgnings data har en prædiktiv kraft for investor adfærd. Et andet bidrag er identifikationen af private og professionelle investorers forskellige brug af Google søgninger, og den høje betydning af aktie symboler i feltet af alle aktie relaterede Google søgninger når man laver prædiktiv modellering af investor adfærd.

Denne PhD afhandling bidrager til den akademiske diskussion om identificering af prædiktive krafter i sociale data, og for den prædiktive modellering af købs- og investor-adfærd. Hoved bidraget er præsentationen af en ny konceptuel filtrerings model for sociale data i kapitel 7, Figure 10: Social Filtering Model., der forklarer hvordan menneskelig adfærd på de forskellige sociale medier og web søgninger er filtreret på forskellig vis. Denne forskel i filtrering bestemmer den potentielle brug af sociale data i en praktisk prædiktiv model bygge anvisning, der også præsenteres i kapitel 8, Table 14, Predictive Modelling Framework.

List of Tables in the Kappa

Table 1. Starting years of social media websites. 20
Table 2: Research questions and key contributions. 26
Table 3: Chapter outline and summary. 35
Table 4: Methods covered by the nine definitions of predictive analytics. 41
Table 5: Differences between explanatory statistical modelling and predictive analysis. 42
Table 6: Predictive models in the literature. 45
Table 7: Social science paradigms. 57
Table 8: Contrasting philosophies of computational social science. 60
Table 9: Scope, methods, insights, and conceptual models in all five papers. 64
Table 10, LazyPredict results for paper I, predicting iPhone sales with Twitter data 91
Table 11, LazyPredict results for paper II, predicting H&M sales with Facebook data 98
Table 12, LazyPredict results for paper III, predicting Mikkeller sales with Google search data 104
Table 13, LazyPredict results for paper V, predicting Apple stock price volatility with Google search data 109
Table 14, Predictive Modelling Framework 129

List of Figures in the Kappa

Figure 1: Sizes of the user bases of the largest social media websites 21
Figure 2: Steps for building an empirical model (predictive or explanatory). 53
Figure 3: Customer infinity model 3.0. 67
Figure 4: Steps in the statistical modelling process. 68
Figure 5: CRISP-DM six-step modelling process diagram. 69
Figure 6: Copenhagen Business School Facebook post. 74
Figure 7: Q4.14 prediction of iPhone sales. 75
Figure 8: Q4.14 prediction of H&M sales from Facebook likes. 76
Figure 9, R-square formula from Scikit Learn 85
Figure 10: Social Filtering Model. 114
Figure 11: Social Filtering Model by Lassen (2023) combined with the social data model by Vatrapu et al. (2016). 128
Figure 12, Digital Maturity Model, by Boston Consulting Group (2021) 135
Figure 13: Awareness, research, decision, and purchase journey model. 139
Figure 14: Google searches for TikTok, Twitter & Instagram 2018–2023. 141

List of abbreviations

SNA	Set network analysis
SSA	Social set analysis
IS	Information systems

Table of contents

Acknowledgements **3**

English Abstract **5**

Dansk Abstrakt **7**

List of Tables in the Kappa **10**

List of Figures in the Kappa **10**

List of abbreviations **10**

Table of contents **11**

1. Introduction **19**

1.1 Motivation 23

1.2 Problem Definition 24

1.3 Research Questions and Key Contributions 25

1.4 Argumentation of the Thesis and Research Papers 29

1.5 Thesis Outline 34

2. Related Work **36**

2.1 Predictive Models Using Social Data 36

2.2 Predictive Modelling vs Explanatory Modelling 41

2.3 Predictive Models assisting researchers 46

3. Empirical Cases and Datasets **50**

4. Research Philosophy and Methodology **53**

4.1 Research Approach 53

4.2 Steps for building an empirical model according to Shmueli and Koppius (2011) 54

4.3 Steps for building a model according to CRISP-DM 54

4.4 Philosophy of Science 56

4.5 Methodological Foundation 59

4.6 Social Network Analysis vs Social Set Analysis 59

4.7 Customer Journey Models 63

5. Design of Predictive Models.....68

Step 1. Shmueli et al (2011) Problem formulation & CRISP-DM Problem Definition.....69

Step 2. Shmueli et al (2011) Data Collection & CRISP-DM Data requirements 70

Step 3. Shmueli et al (2011) Data preparation & CRISP-DM Data preparation 70

Step 4. Shmueli et al (2011) Model specification & CRISP-DM Modelling 70

Step 5. Shmueli et al (2011) Model estimation & CRISP-DM Modelling.....71

Step 6. Shmueli et al (2011) Model evaluation & refinement, CRISP-DM Evaluation.....72

Step 7. Shmueli et al (2011) Results, CRISP-DM Deployment 73

6. iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models 78

6.1 Choice of datasets.....78

6.2 Limitations of linear regression models 79

6.3 Changing from linear models in paper I & II, to a non-linear model in paper III 80

6.4 Changing from statistical models in paper I, II & III, to machine learning models in paper V.....80

6.5 Comparing two machine learning models in paper V.....82

6.6 iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models 83

6.7, LazyPredict 40 models results for paper 1, predicting iPhone sales with Twitter data....85
6.7.1 Bagging and boosting models....86
6.7.2 Analysis....87

6.8, LazyPredict 40 models results for paper II, predicting H&M sales with Facebook data....92
6.8.1 Regularized and generalized linear models (GLM)....92
6.8.2 Analysis....94

6.9, LazyPredict 40 models results for paper III, predicting Mikkeller sales with Google search data....99
6.9.1 Adaboost, regularized and generalized linear models (GLM)....99
6.9.2 Analysis....101

6.10, LazyPredict 40 models results for paper V, predicting Apple stock volatility with Google search data....104
6.10.1 Boosting and bagging, OMP, kNN and Transformed Target models....105
6.10.2 Analysis....106

6.11 Conclusions for the LazyPredict models on iPhone, H&M, Mikkeller and Apple datasets....110

7. Social Filtering Model....111

7.1 Dimensions of social media filtering....115
7.1.a The filters individuals apply to the content they publish....115
7.1.a.1 Online norms for negative and positive emotions on social media....115
7.1.a.2 Social comparisons on social media....118
7.1.b The filter bubble algorithms social media sites uses to personalise content based on users' interest....120
7.1.c The API filters social media sites use for access to their data....121
7.1.c.1 Twitter API....121
7.1.c.2 Facebook API....122
7.1.c.3 Instagram API....123
7.1.c.4 TikTok API....124
7.1.c.5 Google Trends API....125
7.1.c.6 Overall API landscape....125

7.2 Conclusions of the Social Filtering Model....126

8. Predictive Modelling Framework....129

8.1 The steps of the predictive modelling framework.....130

8.2 Guidelines for size of dataset, in the Predictive Modelling Framework....131

8.2 Digital maturity impacts the use of social data 134

9. Findings 136

RQ1: Which social data types can be used to predict consumer purchase behaviours and to what extent does it work for different brand types? 137

RQ2: What, if any, are the explanatory mechanisms for social data based predictive models for consumer purchase behaviours? 138

RQ3: To what extent can social data provide predictors for investor behaviour? 139

RQ4: How can extant social data models be adapted to better inform the predictive models of consumer and investor behaviours? 140

10. Conclusions.....143

10.1 Contributions to the Literature 143

10.2 Managerial Implications.....145

10.3 Limitations.....146

10.4 Future Research Directions 149

References.....150

Appendices.....172

Appendix 1, python LazyPredict code for paper I, predicting iPhone sales with Twitter data.....172

Appendix 2, python LazyPredict code for paper II, predicting H&M sales with Facebook data.....174

Appendix 3, python LazyPredict code for paper III, predicting Mikkeller beer sales with Google searches.....180

Appendix 4, python LazyPredict code for paper V, predicting Apple stock price volatility with Google searches 186

Paper I: Predicting iPhone Sales from iPhone Tweets 189

I.1. Introduction 191

I.2. Related Work 192
I.2.1. Social Data & Business Outcomes: Data Science 192
I.2.2. Social Media Analytics: Information Systems 193

I.3. Theoretical Framework 194
I.3.1. AIDA 195
I.3.2. Hierarchy of Effects (HoE) 196

I.4. Methodology 198
I.4.1. Dataset 198
I.4.2. Quantity of Tweets 198
I.4.3. Quality of Tweets 198
I.4.4. Seasonal Weighting of Tweets 199
I.4.5. Overall Model 200

I.5. Results 201

I.6. Discussion 205
I.6.1. Implications for Organizations 211

I.7. Conclusion 212

I. References 212

Paper II: Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data 218

II.1. Introduction 218

II.2. Towards a theory of social data 219

II.3. Methodology and analytical findings 220

II.4. Discussion 234
II.4.1. Implications for Theory 234
II.4.2. Implications for Practice 235
II.4.3. Limitations 236
II.4.3. Future Work 237

II. References 238

Paper III: Social Media Data as Predictors of Mikkeller Sales? 243

III.1. Introduction 243

III.2. Briefly on the existing literature.....244

III.3. The data and methodology 245

III.3.1 Pre-processing methodology 246

III.3.2 Unobserved Component Models 247

III.3.3 The regression models 248

III.4. Descriptive statistics.....250

III.5. Unobserved components models for the sales series and the social media series. 254

III.6. Results of predictive modeling at the monthly frequency. 257

III.6.1.1 In sample regression results.....258

III.6.1.2 Out-of-sample predictive power? 259

III.6.2. Predicting the irregular component for the sales series by the irregular components for the social media series. Initial investigations. 261

III.7. Summary and conclusion 262

Paper IV: Predictive Analytics with Social Media Data.....264

IV.1. Introduction 264

IV.2. Predictive Models vs. Explanatory Models 265

IV.3. Predictive Modelling of Social Media Data.....267

IV.3.1. The phenomenon of interest and the type of forecasts.....267

IV.3.2.The data 268

IV.3.3. Social media data and pre-processing 269

IV.4. In search of a model equation – theory-based versus data-driven? 269

IV.4.1. Fitting of a predictive model 271

IV.4.2. Evaluation of a predictive model for forecasting purposes.....272

IV.4.3. In-sample evaluation of the model.....272

IV.4.4. Out-of-sample evaluation.....273

IV.4.5. Using a predictive model for forecasting purposes.....273

IV.5. Categorized List of Predictive Models with Social Media Data 274

IV.5.1. Application Domains 274

IV.5.2. Social Media Data Types 278
IV.5.3. Independent and Dependent Variables 278
IV.5.4. Statistical Methods Employed 278
IV.6. An Illustrative Case Study of Predictive Modelling 278
IV.7. Conclusion 281
IV. References 283
Paper V: Google searches linked to Apple stock volatility ups and downs .291
V.1. Introduction 292
V.2. Literature review 294
V.3. The Investor Journey Model 298
V.4. Methodology 299
V.5. Data 301
V.5.1. Google searches during 2015–2020 301
V.5.2. Apple announcement data 307
V.5.3. Weekly return data 309
V.5.4. Formulas 309
V.6. Results and Discussion 311
V.6.1. Diagnostic tests 312
V.7. Forecasting evaluation 314
V.7.1. Autometrics recursive graphs 317
V.7.2. IBM SPSS 26 output 318
V.8. Final model 318
V.9. Conclusions 320
V. References 322
V. Appendix 1. Forecast output from Oxmetrics, September 2019–April 2020, including the COVID-19 pandemic period 326
V. Appendix 2. Forecast output from Oxmetrics, September 2019–February 2020, excluding the COVID-19 pandemic period 327

1. Introduction

The roots of predictive analytics can be traced back to the 1940s when computational models were initially employed by governments. However, simpler forms have existed for thousands of years. For example, around 100 BC, the Antikythera mechanism was a Greek analogue computer that could predict the astronomical positions of celestial bodies decades in advance, as well as eclipses for calendar and astrological purposes.

In sales and marketing research, predictions form a discipline that has historically been based on classical statistical methods, including time series analysis, with a focus on historical sales data as important predictors.

The use of social analytics for predictive models within sales and marketing has been increasing over the past 10 years, especially over the past 5 years. It is cheaper to obtain human behaviour data from social media and web searches compared to classical focus groups and interviews. A negative side of social data used in predictive models is that the reference groups on social media and web searches might not be representative. Bias and limits for topics covered are also among the drawbacks. Social analytics is also offering better data access that can be updated in real-time and is a good supplement to the original data sources.

Social media was born in the 1990s, as the Internet became widely available. The first large social media sites started in 1995 with Classmates, followed by Six Degrees in 1997. Numerous social media sites have sprung up worldwide over the following decades, but only a small number became dominant over time (see Table 1).

Social media site	Starting year
Classmates	1995
Six Degrees	1997
QQ	1999
Ryze	2001
Friendster & Meetup	2002
LinkedIn, hi5, MySpace	2003
Orkut, Flickr, Facebook	2004
Bebo, Yahoo 360°, YouTube, Reddit	2005
Twitter	2006
Tumblr, Vkontakte	2007
Sina Weibo, WhatsApp	2009
Ask.fm, Instagram, Pinterest	2010
WeChat, Snapchat, Google+	2011
Kuaishou	2012
Vine, Telegram	2013
TikTok, Douyin	2016

Table 1. Starting years of social media websites.

Figure 1 below illustrates a visual comparison of the largest social media platforms in 2023, based on their active user count, represented in millions.

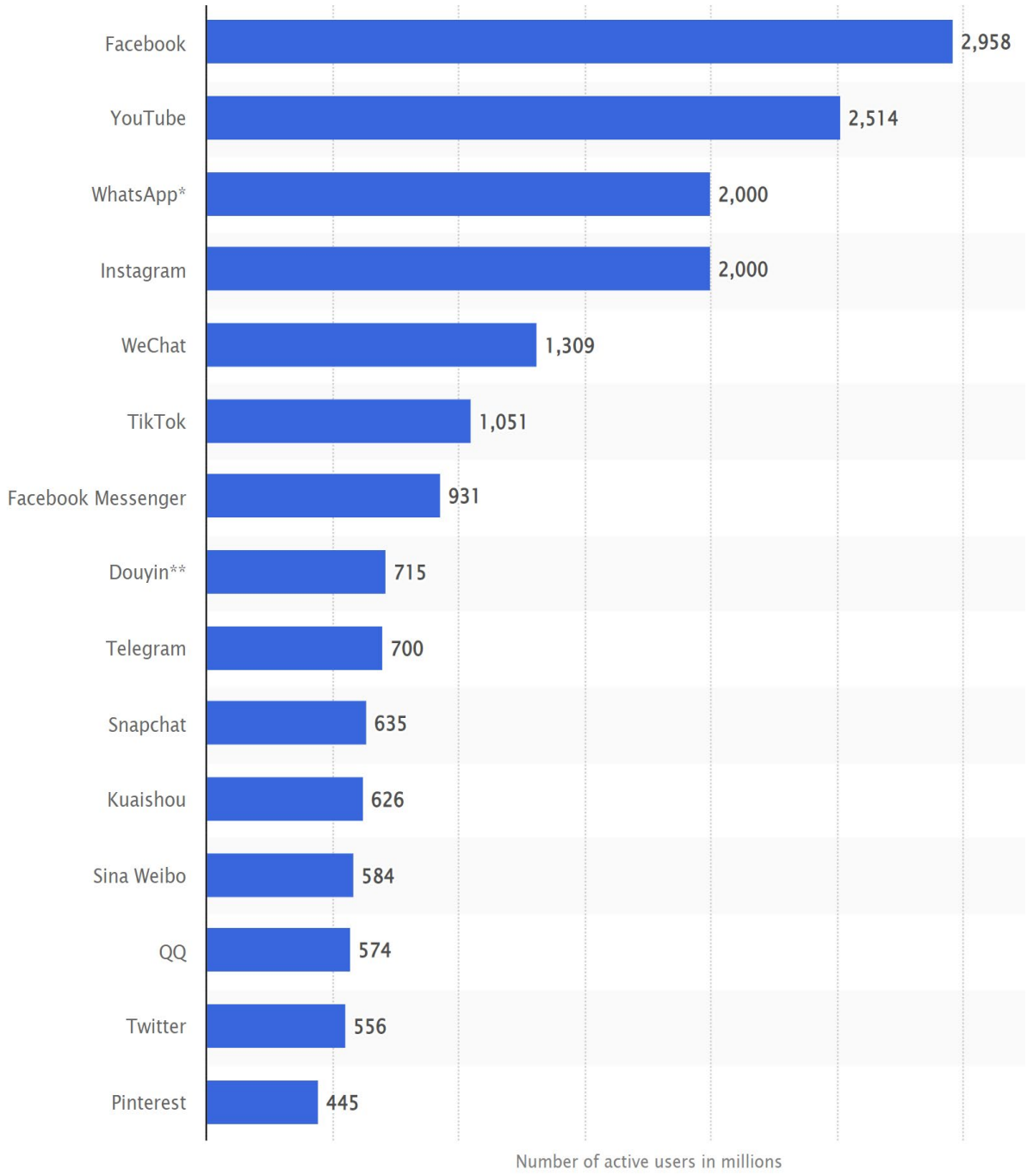


Figure 1: Sizes of the user bases of the largest social media websites

Source: Statista (2023), copyright permission from Statista

Figure 1 shows that Facebook is by far the largest social media site with almost 3 billion users, but TikTok and other sites are growing at a much faster rate than Facebook. Sites like YouTube, Instagram & TikTok are also performing much better than Facebook on engagement. As such, it looks like Facebook will not keep

its top position in the long run.

Social media data were first used for sales forecasting in 2010, when Asur and Huberman (2010) showed how Twitter data can be used to predict Hollywood movie sales. This paper was the starting point for the development of new sales and marketing forecasting models that use social media in conjunction with social data. It also prompted financial markets to realise the predictive power in social media data and, since then, social media data have played an increasingly larger role in forecasting in the areas of sales, marketing, and financial markets. This historical perspective and development also led to the research within this PhD thesis, with social media based forecasting models for human purchase behaviour being adapted to predict human financial behaviour.

Social media data can in many cases be downloaded in limited access via API. More automatic solutions for downloading social media data includes SalesForce's Social Studio, a leading social media analytics tool used in sales, marketing, and financial markets. However, due to its high cost, it is typically utilized by major brands. Affordable alternatives with fewer functionalities include Adobe Social, Zoho Social, Sprinklr, Hearsay Social, Hootsuite, Khoros, Sprout, Falcon, MeltWater, Dataminr, SentiOne, Google Alerts, and Talkwalker. Social media analytics software API solutions often allows access to more social media data compared to free access API, but also comes with a higher price tag.

Stockpulse.com is an example of a social media analytics software aimed at the financial markets and has been in use since 2011. However, there are many competitors to Stockpulse.com, with similar analytical software that collects social data for financial modelling. There are also many startups whose analytics tools use a mix of social media data sources converted into sentiment and emotions as input data, which are then directly used for modelling within financial markets. If

a social media analytics company can show the predictive power of a constructed variable based on social media data, it is often enough to start selling their analytics to financial organisations.

Black Swan is an example of a startup company, utilizing social media and web searches to spot new trends and predict them. In 2011, they started to predict box office sales for Disney and, in 2019, they were serving 16 big brands within consumer goods (Abboud 2019). Black Swan operates in a customer insights market, along with other startup companies such as Signals Analytics and TasteWise. These formerly small startups can add new customer insights for big brands with their big internal marketing and insight departments. Timely access to interview data is also sped up by newer startups such as Zappi and Streetbees and older startups such as SurveyMonkey.

During the research process, I have observed important filtering differences between social media platforms and web searches, but the literature proposes no theoretical or practical model for showing these filtering differences. Cookingham et al. (2015), Gündüz (2017), Fardouly et al. (2018), and Anderson et al. (2018) show that social media provides an unrealistic view of others' lives and affects peoples' identities and mood. The filtered images of other peoples' lives on social media have been researched and documented for more than a decade, but a model showing the difference in filtering among different social media platforms is still lacking.

1.1 Motivation

Predictive models can be highly accurate in forecasting outcomes, but often lack interpretability, making it difficult to understand why they work, see Ribeiro et al (2016) and Guidotti et al (2018). This "black box" problem has motivated me to

develop more explainable models, using conceptual models to explain the predictive power in the predictors.

I have also been motivated to explain the differences in predictive power for different social media and web search data. I was puzzled why there was not a conceptual model for these data differences in the literature, and I was highly motivated to make a new conceptual model. This new model is presented in chapter 7. Social Filtering Model.

I was also motivated to make a practical guide for predictive model building with social media and web search data, as I could not find such a practical guide in the literature in my first papers in this PhD. This own developed practical guide is presented in chapter 8. Predictive Modelling Framework.

1.2 Problem Definition

Filtering of social data is explained in chapter 7. Social Filtering Model, but a short introduction to filtering is given here.

The high-degree filtering of social media data on platforms such as Instagram, Facebook and TikTok happens when these platforms highlight the successes in our lives, while failures are not popular to show. There is a high focus on likes, re-posting and comments. This results in a glossy picture of peoples' lives, which is not representative and limits the use of these data.

Twitter, blogs, and forums are examples of medium-degree filtered social data. People still care about their identities on these platforms, but life failures are allowed here to a higher extent. The system with likes, re-posting and comments, are still creating a filtered picture of of peoples' lives, as many are focused on these social media reactions.

Google searches, and other web searches, are the most un-filtered social data identified in the research of this PhD. When we web search, we do not have focus on likes, re-posting and comments, and we can search for anything we want. The only filter, is the potential worry about an employer watching our web searches, or if we web search next to other people.

This PhD thesis is trying to solve the problem, of giving a better overview of filtering differences of social media and web search data, and what this means for the predictive power and potential uses of these data.

1.3 Research Questions and Key Contributions

Objective 1	**RQ1: Which social data types can be used to predict consumer purchase behaviours and to what extent does it work for different brand types.**
	Paper I demonstrates it is possible to predict iPhone sales using Twitter data, paper II demonstrates H&M sales can be predicted with Facebook data, while paper III demonstrates it not possible to predict Mikkeller beer sales with neither Google nor Facebook data. Research into the reasons for these predictive modelling successes and failure has shown important insights about both the nature and size of social data, thus determining the potential use of social data in predictive models. This is further explained by in the chapter 7. Social Filtering Model and the chapter 8. Predictive Modelling Framework

Objective 2	**RQ2: What, if any, are the explanatory mechanisms for social data based predictive models for consumer purchase behaviours?**
	Customer journey models are used in the first four papers of this PhD thesis, as conceptual models explaining how all product related social data can be placed in one of the phases of a customer journey model. As social data are proxies for the activity in each phase of a customer journey model, the social data acting as proxies for the last two phases of the model and including strong links between human behaviour on social data and purchasing behaviour.
Objective 3	**RQ3: To what extent can social data provide predictors for investor behaviour?**
	Based on the predictive modelling foundations of papers I–IV, a novel investor journey model is developed in paper V for predicting Apple investor behaviour using Google search data as predictors. The main contribution of paper V is the creation of the investor journey model, explaining why the Google search data have predictive power for investor behaviour. Another contribution is the identification of private and professional investors' different use of Google searches and the high importance of stock symbols, among all stock related Google searches for the predictive modelling of investor behaviour.
Objective 4	**RQ4: How can extant social data models be adapted to better inform the predictive models of consumer and investor behaviours?**
	The Social Filtering Model was developed based on the experiences in this PhD and has been adapted into a practical Predictive Modelling Framework for marketing and finance. The two models are presented in chapter 7. Social Filtering Model and chapter 8. Predictive Modelling Framework. The Social Filtering Model shows important filtering differences in the nature of social data from different platforms, which can be in turn used to determine the potential use of social data.

Table 2: Research questions and key contributions.

visualizing that all input data belongs to specific phases of a customer journey model. The main focus has been to generalize how the predictive models in papers I–V can be applied to other brands and stocks. From a practical model building perspective, the type of product, service, or stock being modelled is not the determining factor for model success. Some brands are popular on web searches and social media, while others are not. For example, it is difficult to model insurance or bank revenue, as these services are not popular topics on social media or web searches, even in the case of well-known brands. Therefore, for assessing the potential predictive power of social data for a brand, it is important how popular a brand is on social media and web searches, but also the size of the dataset.

On the one hand, the success of the predictive models for iPhone sales in paper I and H&M sales in paper II proves the predictive power of both popularity and size on social media. On the other hand, the failure of the predictive model for Mikkeller beer sales showed that not only popularity on social media and web searches is important, but also is the size of the brand in general and on social media and web searches. Finally, the success of the predictive model for Apple stock volatility in paper V further proves the predictive power of both popularity and size on social media and web searches.

Another focus of my PhD research has been building predictive models that can provide new logical explanations for the predictive power of social data by using social set analysis (SSA) and the customer and investor journey models. These conceptual models are used for building a novel empirical model in terms of goal definition, data collection, and study design. The use of customer journey models in papers I and II, has led to an important meta contribution. When assessing which phase of the customer journey model an iPhone tweet or H&M Facebook like belonged to, interesting insights came up as follows.

During 2010–2014, more than half a billion iPhones were sold. The potential and actual iPhone customers were a significant part of the Twitter dataset during 2010–2014 in paper I, but a significant part of the tweets containing "iPhone" were from customers who ended up buying competing smartphone brands or who simply had an opinion about the iPhone but no buying intentions. In other words, it would be difficult to precisely place all half billion tweets containing "iPhone" from 2010–2014 within one of the phases of the customer journey model. Despite these problems, the research and work presented in papers I and II yielded a domain-specific classifier for social media texts in the AIDA customer journey's four phases developed by the Center For Business Data Analytics (see http://bda.cbs.dk/). This classifier determines in which phase of the customer decision journey model should a social media text be placed.

This PhD also makes an important contribution to the literature by introducing chapter 7. Social Filtering Model, which identifies important differences in social data, thus determining the potential uses of social data. The Social Filtering Model was forming the basis for the chapter 8. Predictive Modelling Framework, which is the second social data model developed during my PhD.

1.4 Argumentation of the Thesis and Research Papers

Paper I: Predicting iPhone Sales from iPhone Tweets (Lassen et al. 2014)

In the Proceedings of the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, 1–5 September 2014, Ulm, Germany.

Recent research in the field of computational social science have shown how data resulting from the widespread adoption and use of social media channels such as Twitter can be used to predict outcomes such as movie revenues, election winners, localized moods, and epidemic outbreaks. Underlying assumptions for this research stream on predictive analytics are that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product and that the shared digital artefact that is persistent can create social influence. In this paper, we demonstrate how social media data from Twitter can be used to predict the sales of iPhones. Based on a conceptual model of social data consisting of social graph (actors, actions, activities, and artefacts) and social text (topics, keywords, pronouns, and sentiments), we develop and evaluate a linear regression model that transforms iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks. This strong correlation between iPhone tweets and iPhone sales becomes marginally stronger after incorporating sentiments of tweets. We discuss the findings and conclude with implications for predictive analytics with big social data.

Paper II: Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data (Lassen et al. 2016)

In the Proceedings of the 38th Symposium i Anvendt Statistik, Copenhagen Business School, Frederiksberg, Denmark, 25–27 January 2016.

In this conference book chapter, we will advance a theory of social data that distinguishes between constituent dimensions of social graph (i.e., socio-technical affordances of social media networks) and those of social text (i.e., communicative and linguistic properties of social media interactions) as distinct but complementary elements of predictive big social data analytics. Additionally, to illustrate the validity and applicability of our proposed theory, we adhered to the schematic steps advocated by Shmueli and Koppius (2011) in building empirical predictive models that blend social graph analysis with social text analysis to: (1) compute correlations between social data from multiple social media platforms (i.e., Facebook and Twitter) and the financial performance (i.e., quarterly revenues) of corporate entities (i.e., iPhones and H&M), as well as; (2) make predictions about the future performance of these corporate entities. In doing so, we endeavor to provide an answer to the following research question: How can big social data analytics be utilized to predict business performance?

This paper comprises four sections, inclusive of this introduction. In Section 2, we construct our theory of social data by extending Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Section 3 outlines our methodological strategy for extracting and analyzing big social data to build empirical predictive models of business performance. Results from analyzing these empirical predictive models are also reported in Section 3. The last section, Section 4, summarizes the: (1) implications of this study to both theory and practice; (2) insights to be gleaned

towards informing the application of predictive analytics to big social data; (3) possible limitations in the interpretation of our empirical findings, and; (4) probable avenues for future research.

Paper III: Social Media Data as Predictors of Mikkeller Sales? (Lassen et al. 2017)

In the Proceedings of the 39th Symposium i Anvendt Statistik, University of Southern Denmark, Odense, Denmark, 23–24 January 2017.

In recent years, social media data such as Twitter, Facebook and Google Trends data have proven promising as predictors for measures of economic outcomes of private firms. The main advantage of using social media data as predictors lies in the speed with which such data can be extracted and employed in the forecasting process. Once a firm has learned how to collect and pre-process their social media data, the information is available almost in real time and this implies that such data in combination with a good predictive model will provide a very useful tool for the management of the firm.

When working with social media data the concept of 'Big data' often comes to people's minds. In our case this is only partly true: we do work with large amounts of social media data, but once they have been pre-processed, we end up as many studies in the literature using quite simple dynamic regression models based on rather few time series observations. Hence the whole distinction between 'tall', 'fat' and 'huge' data as suggested in Doornik & Hendry (2014) becomes of less relevance. Ideally, if we were able to get economic performance data for a firm at a high frequency like the daily frequency, we would move closer to a situation where a more automatic model selection procedure would be relevant.

The novelty of the present paper is a predictive model for the total sales of Mikkeller using data at a monthly level. With these data we are allowed to be

more precise when it comes to specification of the lag-structure in the dynamic regression model. Also we look into the importance of the data-preparatory work - in our case an unobserved component filtering of the data prior to regression modeling - on the social data proves to be for the final model and finally, we investigate the predictive power of types of social media data that have not been used as predictors before for a brewing company: Google shopping and YouTube data.

Paper IV: Predictive Analytics with Social Media Data (Lassen et al. 2017)
In SAGE Handbook of Social Media Research Methods, chapter 20, pages 328–341, 1st edition, SAGE Publications.

This book chapter provides an overview of the extant literature on predictive analytics with social media data. First, we discuss the difference between predictive vs. explanatory models and the scientific purposes for and advantages of predictive models. Second, we present and discuss the foundational statistical issues in predictive modelling in general with an emphasis on social media data. Third, we present a selection of papers on predictive analytics with social media data and categorize them based on the application domain, social media platform (Facebook, Twitter, etc.), independent and dependent variables involved, and the statistical methods and techniques employed. Fourth and last, we offer some reflections on predictive analytics with social media data.

The authoring team – Niels Buus Lassen, Lisbeth la Cour, and Ravi Vatrapu – were invited to update the chapter on predictive analytics for the second edition of The SAGE Handbook of Social Media Research Method, which was published in 2022.

Paper V: Google searches linked to Apple stock volatility ups and downs (Lassen 2022)

Published in the Conference book of the 43rd Symposium i Anvendt Statistik, Denmark, Axelborg, Copenhagen, Denmark, was presented 29. August 2022 at Axelborg (single-author paper).

Recent studies on how social media and news data are linked to stock price volatility, show that an increase on Twitter is linked to higher volatility, whilst an increase on news media is linked to lower volatility in the following month. This article demonstrates how Google searches are linked to weekly changes in Apple stock volatility. It shows the effects of behavior of private and professional investors on Google searches and how this behavior links to the Apple stock volatility. The paper uses the Customer Journey Mindset from sales modelling to construct a novel "Investor Journey" model, which maps Google searches to investor behavior, currently missing in the literature. Subsequently, the paper summarizes the main findings in this field and outlines future challenges in this research

1.5 Thesis Outline

This doctoral dissertation is composed of ten chapters, coupled with a compilation of five independent scholarly articles, arranged in a sequential manner. Each paper stands on its own and can be comprehensively understood independently, yet their collective contributions offer a unified solution to the research questions previously stated. The initial chapter provides an overview of the research area, while the structure of the remaining chapters is depicted in Table 3.

Chapter		**What does this chapter address?**
Chapter 1: Introduction		Summary of the scope of the research
Chapter 2: Related work	-	Discussion of related work in the domain of predictive models using social data
Chapter 3: Empirical cases and datasets		Presentation of the empirical cases and the datasets.
Chapter 4: Research philosophy and methodology		Presentation of the research philosophy and methodology
Chapter 5: Demonstration of the performance of the predictive models		Demonstrations of the predictive accuracy into the future, and out-of-sample, of the models developed in this PhD
Chapter 6: iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models	-	New analysis for paper I, II, III and V presented, where 40 new models are tested on the iPhone, H&M, Mikkeller and Apple datasets.

Chapter 7: Social Filtering Model	-	Model describing the differences in data filtering for different social media and web searches.
Chapter 8: Predictive Modelling Framework		Practical model for suggesting which social data are relevant for different domains of modelling
Chapter 9: Findings		Presentation of research questions, and how they were addressed and answered in this PhD
Chapter 10: Conclusions		Contributions to the literature, managerial implications, limitations, and future research directions

Table 3: Chapter outline and summary.

2. Related Work

Kalampokis et al. (2013) reviewed 52 articles on predictive models published during 2004–2013 that predicted real world outcomes using social media data as input data. Further, Rousidis et al. (2019), reviewed 40 articles with predictive models published during 2015–2019 that also used social media data as input data.

However, the review of these many articles on predictive models using social media data as input does not provide any logical explanation for why social media data has predictive power. This PhD project addresses this gap in the field of predictive modelling research and examines why social data have predictive power for sales and financial markets. Specifically, Paper I focuses on an iPhone sales prediction model, being the first academic paper to explain why social media data have predictive power for sales (see e.g. Voortman 2015, p.15).

One of the most notable social-data-based predictive models on modelling stock price volatility using Google Trends is that of Preis et al. (2013). They suggest that 'Google Trends data and stock market data may reflect two subsequent stages in the decision making process of investors', which makes Preis et al. (2013, p. 5) is one of the few academic articles presenting a logical explanation for why social data such as Google searches can predict stock price volatility. Their framework is comparable to the conceptual framework developed in paper V, titled 'Google searches linked to Apple stock volatility ups and downs', in which a detailed investor journey model and a logical explanation of the predictive power of Google searches are further developed.

2.1 Predictive Models Using Social Data

This chapter presents the theoretical foundations of this PhD thesis by reviewing the literature on predictive model research using social data and defining a

predictive model within the context of my PhD research. To this end, I review here the existing methods in predictive modelling using social data.

Before proposing a prediction model, it is relevant to specify the differences between forecasting and prediction models. Forecasting is a sub-discipline of prediction, where predictions about the future are made based on time-series data. Therefore, the difference between prediction and forecasting is the use of data with a time dimension. Prediction models using time-series data can also be called forecasting models. Therefore, all forecasting models can be called prediction models, but prediction models cannot be called forecasting models if they are based on data without a time dimension. Numerous definitions of prediction models exist in scholarly works, and a subset of these will be examined in the following discussion.

A practical definition was given by Shmueli and Koppius (2011, p. 3): 'A predictive model is any method that produces predictions, regardless of its underlying approach: Bayesian or frequentist, parametric or nonparametric, data mining algorithm or statistical model etc.'.

Predictive models are also often defined as part of predictive analytics. One definition of predictive analytics comes from one of the leaders in statistical and machine learning software, the SAS Institute (2020), and also includes machine learning models: 'Predictive analytics is the use of data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data'.

A third definition of predictive models comes from one of the leaders in evaluating statistical and machine learning software solutions, Gartner Inc. (2020), who define predictive modelling as 'a commonly used statistical technique to predict future

behaviours. Predictive modelling solutions are a form of data-mining technology that works by analysing historical and current data and generating a model to help predict future outcomes. In predictive modelling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available'.

However, many of the definitions of predictive modelling mostly cover statistical models and do not reference machine learning models, which makes them outdated. For instance, Gartner Inc. do not include machine learning in their definition, which is surprising coming from one of the leading analysts of statistical and machine learning software solutions.

The fourth definition comes from another leader in this field, IBM (2020): 'Predictive analytics is the use of advanced analytic techniques that leverage historical data to uncover real-time insights and to predict future events. The use of predictive analytics is a key milestone on your analytics journey — a point of confluence where classical statistical analysis meets the new world of artificial intelligence (AI)'. It is remarkable that IBM includes AI in their definition but not machine learning in general, which AI modelling is part of. An explanation can be that IBM are branding themselves as one of the leaders within AI, so their definition is more of a branding statement, rather than a precise definition of predictive analytics, which should cover both statistics and machine learning.

In short, the SAS Institute (2020) is more precise in their definition of predictive analytics compared to IBM.

The branding and marketing agendas affecting the definition of predictive analytics come from the battle over market shares. In practice, predictive modelling covering both statistics and machine learning is often carried out using Python, which is currently the main leading data science programming language. Further, both the

SAS Institute and IBM are continually losing market shares to free data science programming languages such as R and Python. A predictive model custom built using a programming language such as Python gives more insights into predictive mechanisms compared to software solutions developed by, for example, SAS Institute or IBM, which always include a degree of black boxing.

The practical leaders within predictive analytics can be identified at Kaggle data competitions, where big tech and similar organisations are very visible as the top 10. Some of these practical leaders, such as Google and Facebook, base their analytics on Python. However, their definitions of predictive analytics are not discussed here, as they are rather marketing statements than scientific definitions.

Microsoft is offering predictive analytics solutions through their Azure and Power BI software platforms, which are comparable to the similar solutions developed by SAS Institute and IBM. Microsoft has the same black box problems as SAS Institute and IBM and is continually losing market share to free and open solutions such as R and Python. The definition of predictive analytics from Microsoft is also more marketing-based than scientific and is, thus, not reviewed here.

The fifth definition of predictive analytics comes from Teradata (2020), which is an analytical software company: ‘Predictive analytics refers to the analysis of big data to make predictions and determine the likelihood of future outcomes, trends or events. In business, it can be used to model various scenarios for how customers react to new product offerings or promotions and how the supply chain might be affected by extreme weather patterns or demand spikes. Predictive analytics may involve various statistical techniques, such as modelling, machine learning, and data mining’. Teradata was formed in 1979 in Brentwood, California, as a collaboration between researchers at Caltech and Citibank's advanced technology group, so the predictive analytics background of this group is comparable to that

of the SAS Institute. Their definition completely covers the current scope of predictive analytics with statistics, machine learning, and data mining.

The sixth definition of predictive analytics comes from (Fayyad et al., 1996). "Predictive analytics is the process of using data mining, machine learning algorithms, and predictive models to identify the likelihood of future outcomes based on historical data"

The seventh definition of predictive analytics comes from (Wu & Zhang, 2014). "Predictive analytics refers to the use of statistical and machine learning techniques to analyze historical data in order to make predictions about future events or trends"

The eigths definition of predictive analytics comes from (Gupta & Bhatia, 2014). "Predictive analytics is the branch of analytics that deals with the extraction of information from data and the use of that information to predict future trends and behavior patterns"

The ninths definition of predictive analytics comes from (Cios et al., 2007). "Predictive analytics is a set of statistical and machine learning techniques that use historical data to identify patterns and make predictions about future events"

These nine definitions differ mainly in which methods from statistics, machine learning, and data mining they include. The reasons for these differences mainly stem from the competition over market share, but some definitions are simply outdated. There is an ongoing competition between statistics and machine learning, where both sides advocating their solutions to data science domains such as predictive analytics. This conflict is one of the main explanations for the many varying definitions of predictive analytics, but cultural beliefs also influence them (e.g. Koehrsen 2019; Bzdok et al. 2018). The practical truth about this conflict is that both sides are needed. In thorough testing environments for predictive models

with big datasets, both statistical and machine learning models are tested on the same datasets and compared against each other. The Danish National Bank published a paper (Christoffersen et al. 2018) showing how machine learning models compete against statistical models on modelling financial distress.

Therefore, in this PhD thesis, I advocate for a definition of predictive analytics that includes statistics, machine learning, and data mining, as covered by Teradata, Fayyad et al., (1996) and Wu & Zhang, (2014).

Definitions of Predictive Analytics									
	Fayyad et al.	Cios et al	Wu & Zhang	Gupta & Bhatia	Shmueli & Koppius	SAS Institute	Gartner Inc.	IBM	Teradata
	1996	2007	2014	2014	2011	2020	2020	2020	2020
Statistics	+	+	+	+	+	+	+	+	+
Machine Learning	+	+	+	+		+		Only AI	+
Data Mining	+	+	+	+	+		+		+

Table 4: Methods covered by the nine definitions of predictive analytics.

2.2 Predictive Modelling vs Explanatory Modelling

The main differences between explanatory and predictive models are shown in the following table.

Step	Explanatory	Predictive
Analysis Goal	Explanatory statistical models are used for testing causal hypotheses.	Predictive models are used for predicting new observations and assessing predictability levels.
Variables of Interest	Operationalized variables are used only as instruments to study the underlying conceptual constructs and the relationships between them.	The observed, measurable variables are the focus.
Model Building Optimized Function	In explanatory modeling the focus is on minimizing model bias. Main risks are type I and II errors.	In predictive modeling the focus is on minimizing the combined bias and variance. The main risk is over-fitting.
Model Building Constraints	Empirical model must be interpretable, must support statistical testing of the hypotheses of interest, must adhere to theoretical model (e.g., in terms of form, variables, specification).	Must use variables that are available at time of model deployment.
Model Evaluation	Explanatory power is measured by strength-of-fit measures and tests (e.g., R^2 and statistical significance of coefficients).	Predictive power is measured by accuracy of out-of-sample predictions.

Table 5: Differences between explanatory statistical modelling and predictive analysis.

Retabulated from: Shmueli and Koppius (2011).

The building block of predictive models is using a train/test split on the dataset, and then utilizing the test part as out-of-sample for testing the accuracy of the predictive model. The most widely used train/test split is 80/20, where 80% of the dataset is used for training and building the predictive model and 20% for out-of-sample testing. The explanatory models trained and built on 100% of the dataset can be tested as predictive models with an 80/20 or other train/test split.

Historically, explanatory modelling has been emphasized more than predictive modelling due to several reasons. Some of these reasons include the objectives of scientific inquiry, the nature of scientific theories, and the historical context of scientific development.

See Kuhn, T.S. (1962), Merton, R.K. (1973), Popper, K. (1959) and Porter, T.M. (1986).

Shmueli and Koppius (2010) illustrated the lack of predictive modeling in the field of Information Systems. After examining the 1072 articles published between 1990 and 2006 in the highly esteemed journals, Information Systems Research and MIS Quarterly, it was discovered that merely 52 of these empirical studies contained

predictive assertions. Moreover, among these, only seven conducted appropriate predictive modeling or testing.

Since Shmueli and Koppius's (2010) study, the field of Information Systems (IS) has witnessed significant developments in predictive modelling practices. The growing importance of data-driven decision-making, advancements in machine learning techniques, and the increasing availability of large-scale data have contributed to the adoption and enhancement of predictive modelling in the IS domain.

See fx Chen et al (2012), Agarwal et al (2014) and Bapna et al (2020).

Despite the historical emphasis on explanatory modelling in many fields, recent advancements in data science, machine learning, and artificial intelligence have led to an increased interest in predictive modelling. As the value of accurate predictions in various fields becomes more apparent, it is likely that the focus on predictive modelling will continue to grow.

The growth of predictive modelling relative to explanatory modelling can be attributed to several factors, such as advancements in technology, the availability of large datasets, and the increasing demand for accurate predictions in various industries.

See Hastie et al (2009), Davenport et al (2007), Varian, H.R. (2014) and Jordan et al (2015).

While predictive modelling is experiencing significant growth, it is important to note that explanatory modelling remains a vital component of scientific research. The two types of modelling often complement each other, with explanatory models providing insights into underlying causal relationships, and predictive models

leveraging those insights to make accurate forecasts.

Some of the leading predictive modelling environments in the industry are companies such as Open AI, Microsoft, Facebook, Amazon, Google and Apple.

The early adoption of machine learning by big tech and other tech has also led to machine learning excellence, which can create serious ethical problems, in the form of dark patterns, that tech companies use to manipulate users (Financial Times 2021).

Paper IV provides an overview of the academic predictive models using social data. In the first version of paper IV (Lassen et al. 2017), 38 predictive model articles were reviewed. Two thirds of the articles were using regression and other statistical models (among which linear regression represent around half of all articles), and only one third were using machine learning models. This thus provides an objective image of the strength of regression models, but also of the more advanced machine learning models starting to be used more widely by academic researchers. In the second version of the same paper, published in 2022 in The SAGE Handbook of Social Media Research Methods, machine learning models constituted almost half of the examined 55 predictive models compared to 4 years before, when only one third of the 38 examined predictive models were machine learning ones. This clearly shows the development of machine learning models for predictive modelling from 2017 to 2021.

Based on the review work in the two versions of paper IV, predictive models using social data are listed and analysed in Table 6.

Predictive models reviewed		
	38 articles 2008–2015	55 articles 2008–2020
Statistical regression models	63%	47%
Other statistical models	11%	7%
Statistical models versus machine learning	5%	20%
Machine learning	21%	26%

Table 6: Predictive models in the literature.

This illustrates the development of predictive models using social data, with machine learning taking up an increasing share of these models.

The Makridakis Forecasting Competitions started in 1982 (Makridakis et al. 1982), and in the fourth M4 competition in 2018 with 100.000 timeseries and 61 forecasting methods (Makridakis et al. 2018), machine learning was, for the first time, part of the winning forecasting hybrid method, combining both statistic and machine learning features. In the fifth M5 competition in 2020 with 42.000 timeseries from Walmart and 5.500 participating teams at Kaggle.com, all 50 top-performing methods used machine learning (Makridakis et al. 2020). The conclusion of the Makridakis competitions is that hybrid approaches and a combination of methods is the way forward to improve forecasting accuracy and make forecasting more valuable.

Although the Makridakis Forecasting Competitions do not include social data, they still confirm the growing importance and use of machine learning within forecasting in general.

The growing use of machine learning within predictive modelling and the

experiences from the Makridakis competitions mean that machine learning is enabling more models using more data. Machine learning can often lead to better predictive models, either in hybrid approaches or in combination with other methods. Therefore, when the classical statistical approach to predictive modelling does not find enough predictive power in independent variables, machine learning can help predictive modelling succeed even for datasets not compatible with classical statistical methods.

2.3 Predictive Models assisting researchers

Shmueli and Koppius (2011) defined six roles under which predictive models can assist researchers, namely:

(i) generating new theory;

(ii) developing measures;

(iii) comparing competing theories;

(iv) improving existing models;

(v) assessing relevance; and

(vi) assessing predictability.

The descriptions of these roles and how they relate to the predictive models developed in this thesis are provided below to better explain the scope of this PhD project.

Role 1 Generating new theory

By testing predictive models on complex large datasets with many features, new theories will naturally evolve. Shmueli and Koppius (2011) give three such examples. The first is the development of a new auction model based on price velocity and the acceleration from the auction start until the time of the prediction (Jank and Shmueli 2010). The second example is given by Stern et al. (2004), who used predictive analytics to identify the factors affecting broadband adoption by Australian households, resulting in the discovery of a new construct, called 'technophilia'. The third example is the work of Wang et al. (2008), who studied the relationship between how firms disclose security risk factors over a certain period and their subsequent breach announcements. Specifically, by using predictive analytics with textual data, the textual content of security risk factors was found to be a good predictor of future breaches, thus shedding light on a relatively unexplored research area. In my research, a model linking Facebook emotions to the Net Promoter Score was identified by testing various data links to the Net Promoter Score. This model could be further developed based on additional data.

The Net Promoter Score model was presented by me at the conference 'Danish Loyalty Clubs' on 29 January 2016 at Copenhagen Business School (please refer to p. 70–72 of Danish Loyalty Clubs conference 2016).

Role 2 Developing measures

New variables are often developed through experimentation but also when existing variables do not have enough predictive power. However, the predictive power of a variable can be improved by transforming a measurement (e.g. converting the measurement per day into occurrences per week). Other methods of transformation

include the log and square transformations or combining variables from different sources into a constructed variable. In the predictive models presented in this PhD thesis, different methods of transforming variables are used, including log and squared transformations, experimenting with different time periods of social media mentions of brands, and using Google searches indexed against each other.

Role 3 Comparing competing theories

Predictive models can evaluate competing theories by comparing their predictive accuracies, whereas explanatory models are not as suitable for this scope. Predictive models can thus identify the theories with better predictive power, and my research is comparing two theories – SSA and SNA – finding that SSA performs better as a predictive model when using big datasets of social-data-based on Google searches.

Role 4 Improving existing models

Explanatory models are not always able to capture complex underlying patterns and relationships. However, by identifying the input variables with predictive power, patterns and relationships are better defined compared to only using input variables that explain historical patterns. In this thesis, papers I–IV identify new patterns and relationships between brand mentions on social media and brand Google searches regarding sales. Paper V further identifies novel relationships between stock-related Google searches and stock volatility.

Role 5 Assessing relevance

If a theoretical explanation can be used to build a predictive model with good predictive accuracy, it is also a good assessment of the relevance of a provided explanation. Conversely, if the predictive accuracy is poor, the explanation is likely to be less relevant. In papers III and V, brand mentions and brand-related Google searches have been identified as proxies for brand attention and activity in all phases of both the customer and investor journey models.

Role 6 Assessing predictability

If new models, variables, and data sources do not improve predictive accuracy during testing, then the predictive accuracy level of existing models may prove a good benchmark.

3. Empirical Cases and Datasets

Paper I: The aim was to conduct predictive modelling of iPhone sales using Twitter data. Monthly and quarterly Twitter iPhone data for 2007–2014 were selected along with various time lags for the same time periods from TopsyPro. The download process was done through 14-day free trial accounts on Topsy.com from October 2013 to June 2014. Apple Inc. bought Topsy Labs, Inc. in December 2013 for USD 200 million, and shut down TopsyPro Analytics and Topsy.com on 16 December 2015 (AppleInsider 2015). The 14-day free trial account possibility also stopped in mid-2014, which is why no more data could be downloaded for the iPhone predictive model in paper I. Other Twitter data were explored after June 2014, but they were simply too expensive to maintain the model. The data were further prepared in Excel for importing into SAS 9.4.

Dataset 1, Paper I:

Company	Data Source	Time Period	Size of Dataset
Apple[1]	Twitter	2007 → October 12, 2014	500 million+ tweets containing "iPhone"

Paper II: The aim was to conduct predictive modelling of H&M sales using Facebook data. Facebook data from the global H&M Facebook page over 2009–2014 were downloaded using the SODATO Facebook scraping software tool. The download format was CSV, data were prepared further in Excel to match the financial quarters of H&M, and various time lags were used for testing. Data were further prepared using Excel for importing into SAS 9.4.

Dataset 2, Paper II:

Company	Data Source	Time Period	Size of Dataset
H & M[2]	Facebook	January 01, 2009 → October 12, 2014	~15 million Facebook events

Paper III: The aim was to do predictive modelling of Mikkeller beer sales using Google data. Google search, YouTube, and Google Shopping Mikkeller data for 2014–2016 were downloaded from Google Trends in CSV format and further prepared in Excel for import into SAS 9.4. A second version of paper III was submitted in 2018 to the same conference, where the Google data were replaced with Facebook data from the SODATO Facebook scraping software tool.

Dataset 3, Paper III:

Company	Data Source	Time Period	Size of Dataset
Mikkeller	Sales Data	January 2014 - September 2016	33 Monthly Datalines *
	Google Searches	January 2014 - September 2016	33 Monthly Datalines *
	YouTube searches	January 2014 - September 2016	33 Monthly Datalines *
	Google Shopping	January 2014 - September 2016	33 Monthly Datalines *

* Thousands of searches for each monthly search index.

Paper IV: The aim was to review the field of predictive models using social data. This is a review article of the datasets of 38 predictive models using social data, and also considered the iPhone & H&M model examples (dataset 1 and 2).

Paper V: The aim was to conduct predictive modelling of Apple stock volatility using Google search data. Google search Apple related-data for 2015–2020 were downloaded from Google Trends in CSV format and further prepared in Excel for importing into IBM SPSS Statistics 26 and Oxmetrics 8.10.

Dataset 4, Paper V:

Company	Data Source	Time Period	Size of Dataset
Apple	60 different Google Searches	April 2015 - April 2020	260 Weekly Datalines, *1
	Weekly volatility for Apple stock	April 2015 - April 2020	260 Weekly Datalines, *1

*1: With thousands to millions of Google searches behind the weekly search indexes.

A larger version of the dataset of all tested variables, can be found in paper V, section V.5. Data, page 301.

4. Research Philosophy and Methodology

4.1 Research Approach

This PhD research started with a prediction model for iPhone sales, using Twitter data as a predictor (paper I), based on the customer journey model perspective. The customer journey model approach, together with SSA, were then used to test Facebook and Google search data as predictors for the sales of H&M and Mikkeller Brewery in papers II and III. This thesis also examines how the characteristics of Twitter, Facebook, and Google search data differ and why they work differently as sales predictors. The insights from papers I–III are used to develop a more general model for predicting business outcomes using social data (paper IV). In sum, in the first four papers, the customer journey model and SSA are the main concepts used to build predictive models using social data and were subsequently used as the basis for creating the novel investor journey model that can predict investor behaviour using Google search data in paper V. The steps used to develop the social-data-based predictive models in all five papers follow the method of Shmueli and Koppius (2011) and are described in Figure 2.

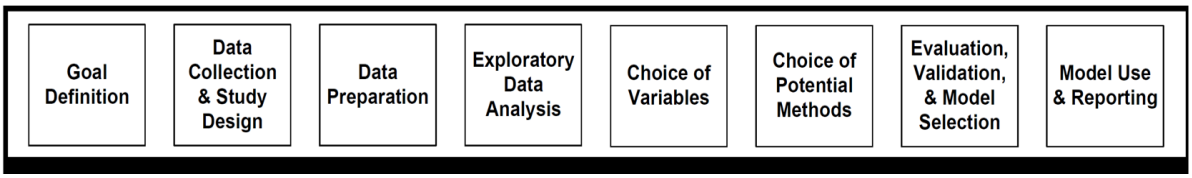


Figure 2: Steps for building an empirical model (predictive or explanatory).

Redrawn from: Shmueli and Koppius (2011).

Shmueli and Koppius (2011) propose a series of steps for building empirical models, which can be either predictive or explanatory. These steps share similarities with the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, which is a cyclical process for building and refining data mining projects, refer to Wirth et al (2000) and Marbán et al (2009). Both approaches

share the same iterative nature and emphasize the importance of continuous improvement and adaptation.

4.2 Steps for building an empirical model according to Shmueli and Koppius (2011)

1. Problem formulation: Define the research question and objectives.
2. Data collection and preparation: Gather and preprocess the data needed to address the research question.
3. Model specification: Specify the structure of the empirical model, including the variables and their relationships.
4. Model estimation: Estimate the model using appropriate statistical techniques.
5. Model evaluation: Assess the model's performance, validity, and reliability.
6. Model refinement: Modify the model based on the evaluation results and iterate through steps 3-5 as necessary.
7. Interpretation and presentation of results: Interpret the results and present them in a clear and concise manner.

The model building steps by Shmueli and Koppius (2011) are very comparable with the six phases of the CRISP-DM model:

4.3 Steps for building a model according to CRISP-DM

1. Business understanding: Define the project objectives and requirements from a business perspective.
2. Data understanding: Collect and explore the data to become familiar with its properties.

3. Data preparation: Preprocess and clean the data to prepare it for modeling.
4. Modeling: Select and apply various modeling techniques to the data.
5. Evaluation: Evaluate the models' performance and choose the best one.
6. Deployment: Implement the chosen model and monitor its performance.

Both the empirical model-building steps by Shmueli and Koppius (2011) and the CRISP-DM model share similar stages, which involve problem formulation, data collection and preparation, model specification, evaluation, and refinement. Both processes are iterative, meaning they may require going back to previous steps, adjusting, and re-evaluating as new information or issues arise (Shmueli & Koppius 2011; Wirth et al (2000) and Marbán et al (2009).)

The circular nature of these processes emphasizes the importance of continuously refining and improving the model to ensure its validity and accuracy. This iterative approach is vital because it allows researchers and practitioners to adapt the model to changing conditions or to incorporate new knowledge and data.

4.4 Philosophy of Science

To reject one paradigm without simultaneously substituting another is to reject science itself.
(Kuhn 1962, page 85)

All computational social science research is based on philosophical assumptions about the world. These assumptions are called paradigms or world views (Crotty 1998) and the most notable ones are listed in below table.

Paradigm	Ontology *What is reality?*	Epistemology *How can I know reality?*	Theoretical Perspective *Which approach do you use to know something?*	Methodology *How do you go about finding out?*	Method *What techniques do you use to find out?*
Positivism	There is a single reality or truth (more realist).	Reality can be measured and hence the focus is on reliable and valid tools to obtain that.	Positivism Post-positivism	Experimental research Survey research	Usually quantitative, could include: Sampling Measurement and scaling Statistical analysis Questionnaire Focus group Interview
Constructivist / Interpretive	There is no single reality or truth. Reality is created by individuals in groups (less realist).	Therefore, reality needs to be interpreted. It is used to discover the underlying meaning of events and activities.	Interpretivism (reality needs to be interpreted) • Phenomenology • Symbolic interactionism • Hermeneutics Critical Inquiry Feminism	Ethnography Grounded Theory Phenomenological research Heuristic inquiry Action Research Discourse Analysis Femenist Standpoint research etc	Usually qualitative, could include: Qualitative interview Observation Participant Non participant Case study Life history Narrative Theme identification etc
Pragmatism	Reality is constantly renegotiated, debated, interpreted in light of its usefulness in new unpredictable situations.	The best method is one that solves problems. Finding out is the means, change is the underlying aim.	Deweyan pragmatism *Research through design*	Mixed methods Design-based research Action research	Combination of any of the above and more, such as data mining expert review, usability testing, physical prototype

Subjectivism	Reality is what we perceive to be real	All knowledge is purely a matter of perspective.	Postmodernism Structuralism Post-structralism	Discourse theory Archaeology Genealogy Deconstruction etc.	Autoethnography Semiotics Literary analysis Pastiche Intertextuality etc.
Critical	Realities are socially constructed entities that are under constant internal influence.	Reality and knowledge is both socially constructed and influenced by power relations from within society	Marxism Queer theory feminism	critical discourse analysis, critical ethnography action research ideology critique	Ideological review Civil actions open-ended interviews, focus groups, open-ended questionnaires, open-ended observations, and journals.

Table 7: Social science paradigms.

Source: Crotty (1998), edited by Salma Patel (2015), who added ontology, and the paradigms pragmatism and critical to the table.

While many researchers have used these paradigms rigidly in their conceptualized models, others are arguing for multi-method approaches, where quantitative and qualitative methods are combined (e.g. Venkatesh et al. 2013; Creswell, 2009). In this PhD thesis, I follow the pragmatism paradigm, because I am a pragmatic researcher. I also bring arguments for this paradigm based on the five dimensions in Table 7.

Ontology. What is reality? Under the pragmatic paradigm, reality is constantly renegotiated, debated, and interpreted in light of its use in new, unpredictable situations. Social media and social data from Google searches could be seen as proxies of human behaviour. Human behaviour and the associated social data are constantly changing. When modelling with social data, reality is dynamic and can only be defined over a short period.

Epistemology. How can I know reality? The pragmatic approach focuses on methods to solve problems. While finding out is the means, change is the

underlying aim. Because this PhD project works with predictive modelling using social media data and social data from Google searches, the reality is constantly changing. Therefore, to define a dynamic reality, a pragmatic and adaptable approach is necessary.

Theoretical perspective. Which approach do you use to find out something? John Dewey (1859–1952) was one the early founders of pragmatism. Dewey's pragmatism was also called 'cultural naturalism' (Dewey 1923). He rejected the dualistic epistemology and metaphysics of modern philosophy and argued for a naturalistic approach that viewed knowledge as coming from an active adaptation of the human organism to its environment. For example, over 2020–2022, we have all been adapting to the COVID-19 pandemic, which is changing the human behaviours related to purchasing, economic decisions, and social interactions. As the significant changes in human behaviour during the pandemic are reflected in our behaviours on social media and those related to Google searches, predictive models built on these data must be redesigned to reflect the new patterns related to human behaviour. Knowledge about human behaviour is constantly changing, which means that the approach to gain knowledge about human behaviour through social data should also be dynamic and re-evaluated frequently.

Research process. How do you go about finding out? Venkatesh et al. (2013) and Creswell (2009) have suggested a combination of quantitative and qualitative methods. This PhD project mainly focuses on quantitative methods mixed with text analytics. Compared to Zaltman's (1996) metaphor elicitation technique (ZMET) interview process, a quantitative big data analytics approach can identify some human behaviours not identified by ZMET interviews, and vice versa. Mixed methods can also be a part of action research, which can result in new insights about human behaviour that lead to social action.

Method. What techniques do you use to find out? Under the pragmatism paradigm, methods can be both quantitative or qualitative, or a combination of both. Further, data mining and machine learning can be used. For this PhD project, I chose quantitative methods covering both statistical and machine learning methods, because it was the most pragmatic choice for big datasets containing the actions of millions of humans.

4.5 Methodological Foundation

This PhD thesis follows a paper-based format, under a pragmatic research paradigm. The research process is structured by the predictive model building steps of Shmueli (2011), also shown in Figure 2: Steps for building an empirical model (predictive or explanatory). The CRISP-DM, see Figure 5: CRISP-DM six-step modelling process diagram., is used for this Kappa and the analysis in chapter 6. iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models.

4.6 Social Network Analysis vs Social Set Analysis

Vatrapu et al. (2016) identified that new computational social science is based on the Social Set Analysis (SSA) framework as an alternative to social network analysis (SNA) for large social media datasets Table 8.

	Social Network Analysis	Social Set Analysis
Basic Premise	There exists a **relation** between social actor A and social actor B	There exists an **association** by actor A with some entity E which can be an actor or an artifact
Social Action	Molecular Relations	Atomic Actions
Unit of Analysis	Dyadic	Monadic, Dyadic & Polyadic
Social Configuration	Networks	Sets
Social Explanation	Structural	Agentic
Mathematics	Graph Theory	Set Theory

Table 8: Contrasting philosophies of computational social science.

Redrawn from: Vatrapu et al. (2016).

Social Network Analysis (SNA) and Social Set Analysis (SSA) are two different approaches used in social research for studying social connections and relationships. While both approaches focus on understanding social structures, they differ in their theoretical assumptions, methods, and analytical techniques.

Social Network Analysis (SNA) is a method used to analyze the structure of social networks. It is based on the assumption that social relationships between individuals can be represented as a network of nodes and edges, where nodes represent individuals and edges represent the social ties that connect them. SNA focuses on understanding the patterns of social connections, such as centrality, density, and clustering, as well as the social processes that shape these patterns. SNA has been widely used in various fields, including sociology, anthropology, psychology, and organizational studies. Refer to Wasserman et al (1994). Social Network Analysis: Methods and Applications. Cambridge University Press.

On the other hand, Social Set Analysis (SSA) is a method used to analyze the structure of social sets. It is based on the assumption that social relationships between individuals can be represented as sets of individuals who share common characteristics or attributes. SSA focuses on understanding the patterns of social sets, such as overlap, complementarity, and substitutability, as well as the social processes that shape these patterns. SSA has been used in various fields, including political science, international relations, and communication studies.

Vatrapu et al. (2016) provide a comprehensive comparison of SNA and SSA. They argue that SNA focuses on studying the structure of social relationships, whereas SSA focuses on studying the structure of social sets. They also note that while SNA is based on the assumption of ties between individuals, SSA is based on the assumption of shared attributes among individuals. Furthermore, they highlight that SNA uses network-based analytical techniques, such as node centrality measures and network visualization, while SSA uses set-based analytical techniques, such as set overlap and set complementarity.

In summary, SNA and SSA are two different approaches used in social research for studying social connections and relationships. While both approaches focus on understanding social structures, they differ in their theoretical assumptions, methods, and analytical techniques.

The **basic premise for** SSA is best explained by an example from Vatrapu et al. (2016). A typical post on F.C. Barcelona's Facebook page in 2016 generated around 100,000 unique likes, 5,000 comments, and 1,000 shares. For such a dataset, SNA would try to map all network connections, from Facebook users' likes to a personal association to one of the players, identity association to the Catalan, political association to the pro-independence parties in Catalonia, brand association to the corporate sponsors etc. Instead, SSA would just presume an

association of the actors giving likes to an entity E that can be an actor or an artefact. In more practical terms, SSA only analyses the actions of humans on social media and avoids the complexity of network relationships. By only focusing on the actions of humans, this method is thus more pragmatic and better suited to complex large social media datasets.

SNA is relevant when the focus is on the relationships between actors in a network, and the structure of the network itself. SNA can help researchers understand patterns of communication, collaboration, and influence among individuals, groups, or organizations in a network. SNA is often used to analyze social media data to identify influential actors, detect communities, and predict information diffusion (Borgatti et al., 2009; Wang et al., 2016). SNA can also be used for modelling reputation (Sabater et al. 2002) and is also suitable for modelling social structures and social influence (Sheedy 2019).

SSA is relevant when the focus is on the attributes of actors and their membership in multiple social groups or sets. SSA can help researchers understand how individuals, groups, or organizations are embedded in multiple social contexts, and how these multiple memberships shape their behaviors, attitudes, or outcomes. SSA is often used to analyze social media data to predict user preferences, behaviors, or attitudes based on their group memberships (Roth et al., 2016; Zeng et al., 2017).

In summary, SNA is more relevant when the focus is on the relationships and network structure, while SSA is more relevant when the focus is on the attributes of actors and their multiple group memberships. Both methods can be used to predict various outcomes with social media data, depending on the research question and the type of data being analyzed.

All datasets in the five papers in this PhD, as presented in chapter 3. Empirical

Cases and Datasets, contain the actions of thousands to millions of humans, so Social Set Analysis, SSA (Vatrapu el at 2016) or set theory (Cantor 1874) are the pragmatic choices for analysing datasets of these sizes. Set theory was proposed in 1874 by Cantor and has been used in many variations in modern mathematics (Ferreirós 2008), while Social Set Analysis, SSA (Vatrapu el at 2016) is also a variant of the original set theory.

Under the pragmatic research philosophy of this PhD thesis, it would not make sense to model the relationships between thousands to millions of humans, which SNA would have required. As such, SNA is relevant for other models of group dynamics as described above, but for modelling sales and stock price volatility with web search and social media data, SSA is the more logical and pragmatic choice.

4.7 Customer Journey Models

In the beginning of my PhD research, a simple customer journey model was chosen as conceptual model in paper I for explaining why Twitter data had predictive power for iPhone sales. This simple model is called the AIDA (attention, interest, desire, action) customer journey model (St. Elmo Lewis 1899) and was chosen for the simplicity and easiness with which it could explain why Twitter data can predict iPhone sales. In the first four papers, the AIDA and hierarchy of effects (HOE) (Lavidge et al. 1961) customer journey models were used to explain the associations and predictive power of social media data for sales.

Table 9 illustrates the use of customer journey models in my papers.

Paper	Scope/RQ	Method	Insight	Conceptual models
Paper I	Can Twitter data predict iPhone sales?	Multiple regression	Social media data can predict sales	AIDA and HOE
Paper II	Can Facebook data predict H&M sales?	Multiple regression	Social media data can predict sales	AIDA and HOE
Paper III	Can Google, YouTube or Facebook data predict Mikkeller sales?	UCM Time Series model	Social media data cannot predict sales, when data too small	AIDA and HOE
Paper IV	38 predictive models using social data reviewed	Statistical and machine learning	Predictive power explained very little	Very few
Paper V	Can Google searches predict Apple stock volatility?	Statistical and machine learning	Web search data can predict stock price volatility	Investor journey model

Table 9: Scope, methods, insights, and conceptual models in all five papers.

All data collected on social media, blogs, forums and web searches for any product or service will belong to one of the customer journey phases. Decision makers should be aware of these marketing data relations for their products and services, as these data can be used for product innovation, product development and spotting new trends before competitors. This is the most important organisational context for all the data collected on social media, blogs, forums and web searches for any product or service. Doing simple topic mining on social media texts about a product or service can often reveal customer preferences for product features, and

also show customer wishes for product features into the future. These customer insights can be used in product innovation, product development and spotting new trends before competitors.

Lemon et al (2016) provides a comprehensive understanding of customer experience throughout the customer journey, emphasizing the importance of considering different touchpoints, including digital channels, to offer better products and services.

Chaffey et al (2019) also show this use of social data in digital marketing strategies, including the use of online data sources to inform decision-making and competitive advantage.

Aslam et al (2020) also show use of sentiment analysis techniques applied to social media data to understand customer opinions and preferences, which can be used to inform product development and innovation.

Gandomi (2015) provide a summary of the key concepts related to big data, encompassing the gathering and examination of information from a variety of online platforms. Highlighting the significance of this data in propelling the innovation of products and services.

Customer journey models are also important methods for explaining the predictive power in social media and web search data as predictors for sales because they help businesses understand the various touchpoints and interactions that customers have with a brand before making a purchase. These models provide valuable insights into customer behavior, which can be used to optimize marketing strategies and ultimately drive sales. (Lemon & Verhoef, 2016).

Customer journey models can also help businesses identify the most influential

touchpoints, allowing them to allocate resources effectively and personalize marketing campaigns for different customer segments (Hübner et al., 2020). This targeted approach can increase the effectiveness of social media and web search data in predicting sales.

By incorporating customer journey models into their data analysis, businesses can improve the accuracy of their sales predictions. These models account for various touchpoints, including social media and web search data, which can increase the predictive power of the analysis (Trainor et al., 2014).

Customer journey models can also help businesses evaluate the effectiveness of different marketing channels, including social media and web search, in driving sales (Edelman & Singer, 2015). By understanding how these channels contribute to the customer journey, businesses can make more informed decisions about their marketing strategies.

In summary, customer journey models are essential methods for understanding and optimizing the role of social media and web search data in predicting sales. By identifying key touchpoints and interactions, businesses can improve their marketing strategies, allocate resources more effectively, and ultimately drive sales.

In paper V, it was the more advanced customer journey model, the customer infinity model (Østergaard et al. 2020), that was the logical basis for developing the investor journey model. The customer infinity model is shown in Figure 3.

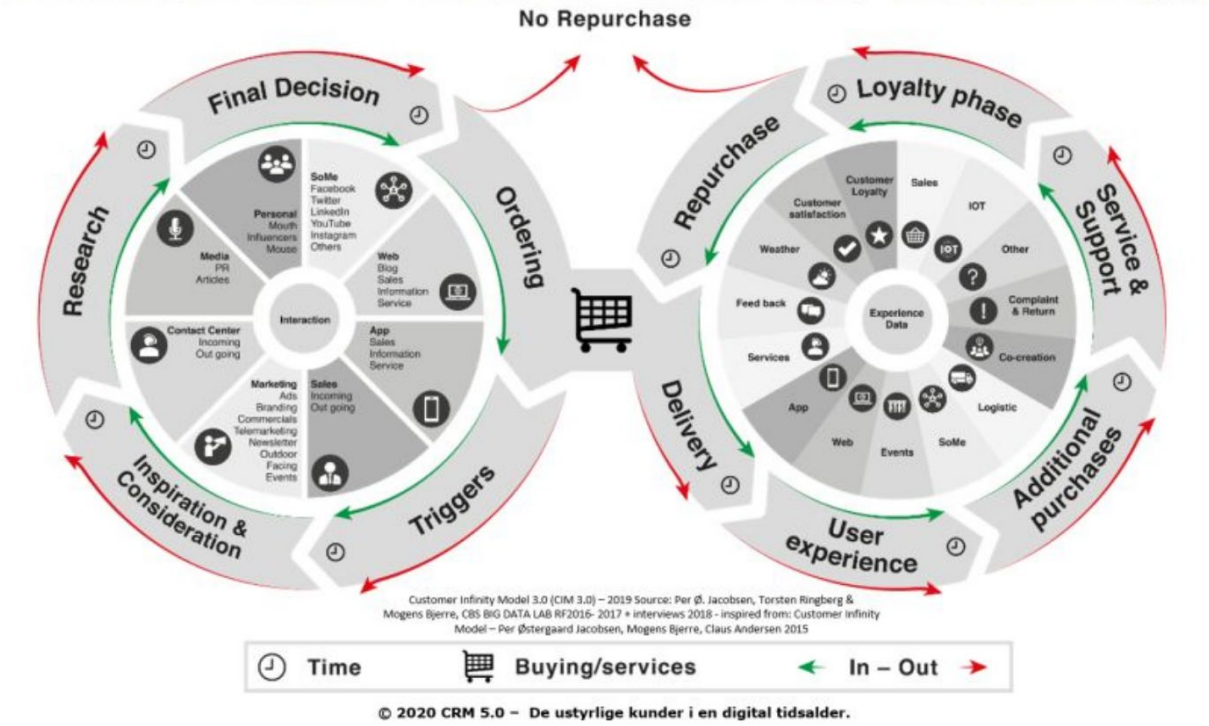


Figure 3: Customer infinity model 3.0.

Source: Østergaard Jacobsen (ed.) 2020, CRM 5.0 - De ustyrlige kunder i en digital tidsalder: Mindset, strategi, ledelse og performance i fremtidens forretningsmodeller. Efficiens, Rungsted Kyst. Copyright permission obtained from Per Østergaard Jacobsen 6. July 2022.

The McKinsey consumer decision journey (McKinsey 2009) and the CBS customer infinity model (Østergaard Jacobsen et al. 2020) are simply more advanced versions of the AIDA and HOE customer journey models, adapted to use digital data from customer relationship management systems, social media, web searches, blogs, forums etc. The customer journey models are very generalizable, focusing on the phases the customers go through before, during, and after a purchase decision. Most industries and companies often need tailor-made versions of the customer journey model adapted to their specific products and services, as well as to industry domain characteristics. Therefore, the customer journey models in all five papers should be considered general examples of conceptual sales models, only meant to show the associations and predictive power of social data for sales.

5. Design of Predictive Models

Shmueli and Koppius (2011) are listing eight steps in their predictive modelling process diagram, which are used as the basis of the research presented within this thesis.

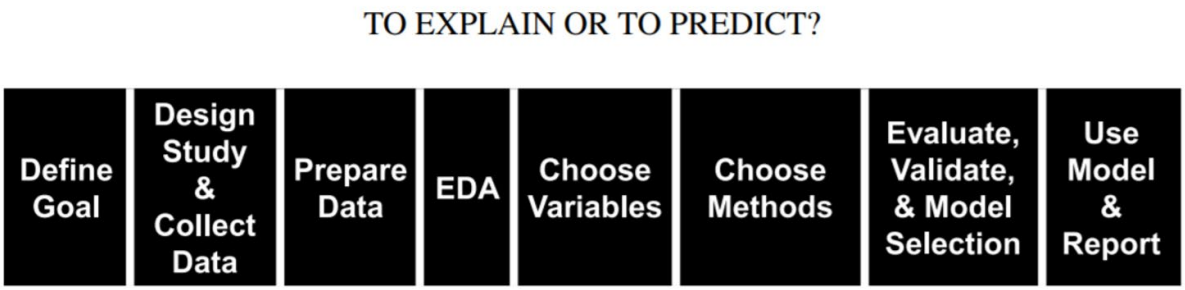


Figure 4: Steps in the statistical modelling process.

Redrawn from: Shmueli and Koppius (2011).

The eight-step model diagram of Shmueli and Koppius (2011) and the six-step model diagram of CRISP-DM (Chapman et al. (2000)) are very similar, as shown in chapter 4.2 Steps for building an empirical model according to Shmueli and Koppius (2011), and chapter 4.3 Steps for building a model according to CRISP-DM.

The CRISP-DM model is shown below.

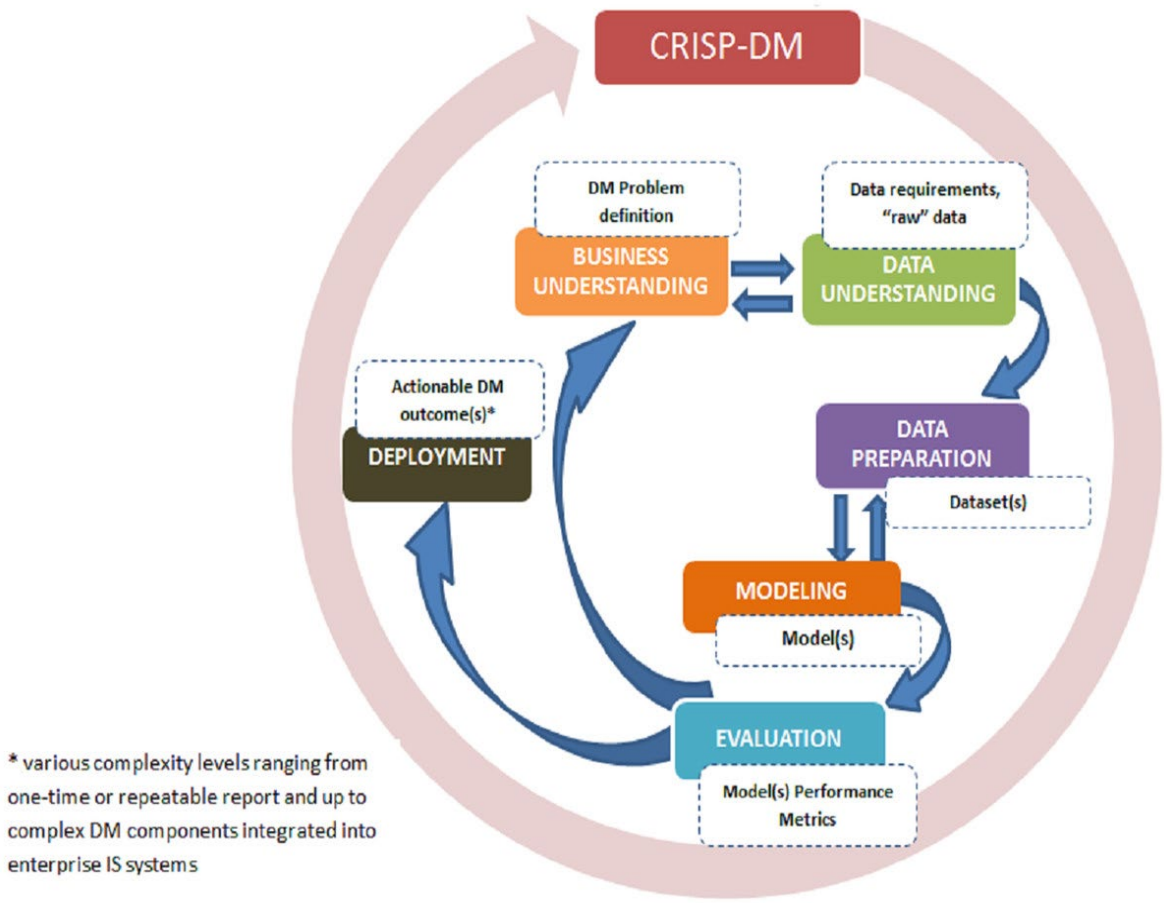


Figure 5: CRISP-DM six-step modelling process diagram.

Adapted from Chapman et al. (2000)

The eight steps of Shmueli and Koppius (2011) and six steps for CRISP-DM modelling processes are compared for the five papers in this PhD below:

Step 1. Shmueli et al (2011) Problem formulation & CRISP-DM Problem Definition

Paper I: Can Twitter data predict smartphone (iPhone) sales?

Paper II: Can Facebook data predict clothing (H&M) sales?

Paper III: Can Google, YouTube, Google Shopping, or Facebook data predict beer (Mikkeller) sales?

Paper IV: What can social media data predict?

Paper V: Can Google search data predict stock (Apple) volatility?

Step 2. Shmueli et al (2011) Data Collection & CRISP-DM Data requirements

Paper I: Which Twitter data are required to predict iPhone sales with Twitter data.

Paper II: Which Facebook data are required to predict H&M sales with Facebook data.

Paper III: Which social data are required to predict Mikkeller sales with Google, YouTube, Google Shopping, and Facebook data.

Paper IV: Explore predictive possibilities with social media data.

Paper V: Which Google search data are required to predict Apple stock volatility with Google search data.

Step 3. Shmueli et al (2011) Data preparation & CRISP-DM Data preparation

For data preparation, please refer to chapter 3. Empirical Cases and Datasets.

Step 4. Shmueli et al (2011) Model specification & CRISP-DM Modelling

Paper I: Twitter iPhone data for 2007–2014 were identified using TopsyPro and selected for testing.

Paper II: Facebook H&M data for 2009–2014 were identified using the SODATO Facebook scraping software tool and selected for testing.

Paper III: Facebook, Google, YouTube, and Google Shopping Mikkeller data were identified using SODATO Facebook scraping software tool and Google Trends, and selected for testing.

Paper IV: Review datasets for 38 predictive models with social data.

Paper V: Google search Apple-related data for 2015–2020 were identified through Google Trends and selected for testing.

Step 5. Shmueli et al (2011) Model estimation & CRISP-DM Modelling

Paper I: The Twitter iPhone data were tested using regression models in SAS 9.4 to model and predict iPhone sales. Time lag transformations were conducted in different combinations to identify the optimal regression model.

Paper II: The Facebook H&M data were tested using regression models in SAS 9.4 to model and predict H&M sales. Time lag transformations were conducted in different combinations to identify the optimal regression model.

Paper III: The Google, YouTube, and Google Shopping Mikkeller data were tested using regression models in SAS 9.4 to model and predict Mikkeller sales. Time lag transformations were conducted in different combinations to identify the optimal regression model.

Paper IV: This is a review article of the datasets for 38 predictive models using social data.

Paper V: Google search Apple-related data over 2015–2020 indexed internally and time lagged before being tested in regression and lasso models using IBM SPSS Statistics 26 and Oxmetrics 8.10 to model and predict Apple stock volatility. Time lag transformations were conducted in different combinations to identify the

optimal regression model.

Step 6. Shmueli et al (2011) Model evaluation & refinement, CRISP-DM Evaluation

Paper I: The AIDA and HOE models were used as conceptual models for explaining both the underlying mechanisms in the model and the predictive power of Twitter data for sales. The over 500 million tweets, considered as quarterly data belong to one of the phases in the AIDA and HOE customer journey models, providing a logical explanation of the predictive power of Twitter data for sales. The logical explanation works if the distribution of Tweets in the different phases of the AIDA and HOE models is relatively stable over time. As such, part of the Twitter data will show a positive relationship with sales, which make the data suitable for using in a simple regression model. Further, the AIDA model provides a logical explanation for why Twitter data need a time lag to work as predictors. Because of the quarterly sales data and limited number of observations, I conducted future predictive testing, as elaborated in Step 7.

Paper II: The AIDA and HOE were used as conceptual models for explaining both the underlying mechanisms in the model and the predictive power of Facebook data. The over 15 million Facebook likes, counted in quarterly periods, belong to one of the phases in the AIDA and HOE customer journey models, thus providing a logical explanation of the predictive power in Facebook data for sales. Because of quarterly sales data and limited number of observations, the predictive testing was conducted for the future, as elaborated in Step 7.

Paper III: The AIDA model was used as conceptual model for explaining the underlying mechanisms in the model. Predictive testing was done on out-of-sample test data, as elaborated in Step 7.

Paper IV: This is a review article presenting the datasets for 38 predictive models with social data, where generalized predictive models for social data were analysed.

Paper V: The logic of the AIDA and HOE models used in Papers I–III was used to develop the investor journey model. This model was then used for explaining both the underlying mechanisms in the model and the predictive power of the Google web search data for Apple stock volatility.

Predictive testing was conducted on out-of-sample test data, as elaborated in Step 7.

Step 7. Shmueli et al (2011) Results, CRISP-DM Deployment

Paper I: This predictive testing was presented at several conferences during 2014–2015. The final version of this article was submitted on 14 June 2014 to the EDOC 2014 Conference. In paper I (available at: https://research.cbs.dk/en/publications/predicting-iphone-sales-from-iphone-tweets), I made a prediction for 37 million iPhones being sold in Q2.14 done before the EDOC 2014 submission deadline on 14 June 2014, which was 1.5 months before Apple released its Q2.14 sales of 35.2 million iPhones at the end of July 2014. The prediction on 37 million iPhones for Q2.14 was 5% over the actual sales on 35.2 million iPhones (Statista 2018), with a prediction accuracy of 95%.

For the actual presentation of the paper at the EDOC 2014 conference on 1 September 2014, the prediction was updated to 36.5 million iPhones for Q3.14. The conference website – Edoc2014.org – is not live anymore, but the prediction graph from the conference presentation is documented by a CBS Facebook post

from 26 September 2014 (see Figure 8) a month before Apple released the Q3.14 iPhone sales at the end of October 2014.

Figure 6: Copenhagen Business School Facebook post.

Source: https://www.facebook.com/CBS.DIGI/photos/professor-ravi-vatrapu-niels-buus-lassen-and-rene-madsen-from-the-computational-/708485655911681/

The prediction of 36.5 million iPhones for Q3.14 was 7% under the actual sales of 39.2 million iPhones (Statista 2018), having a prediction accuracy of 93%.

For ICCSS 2015, a prediction of 68 million sold iPhones for Q4.14 was made by the ICCSS deadline of 7 December 2014. This prediction was 8.7% under the actual sales of 74.47 million iPhones (Statista 2018), for a prediction accuracy of 91.3%.

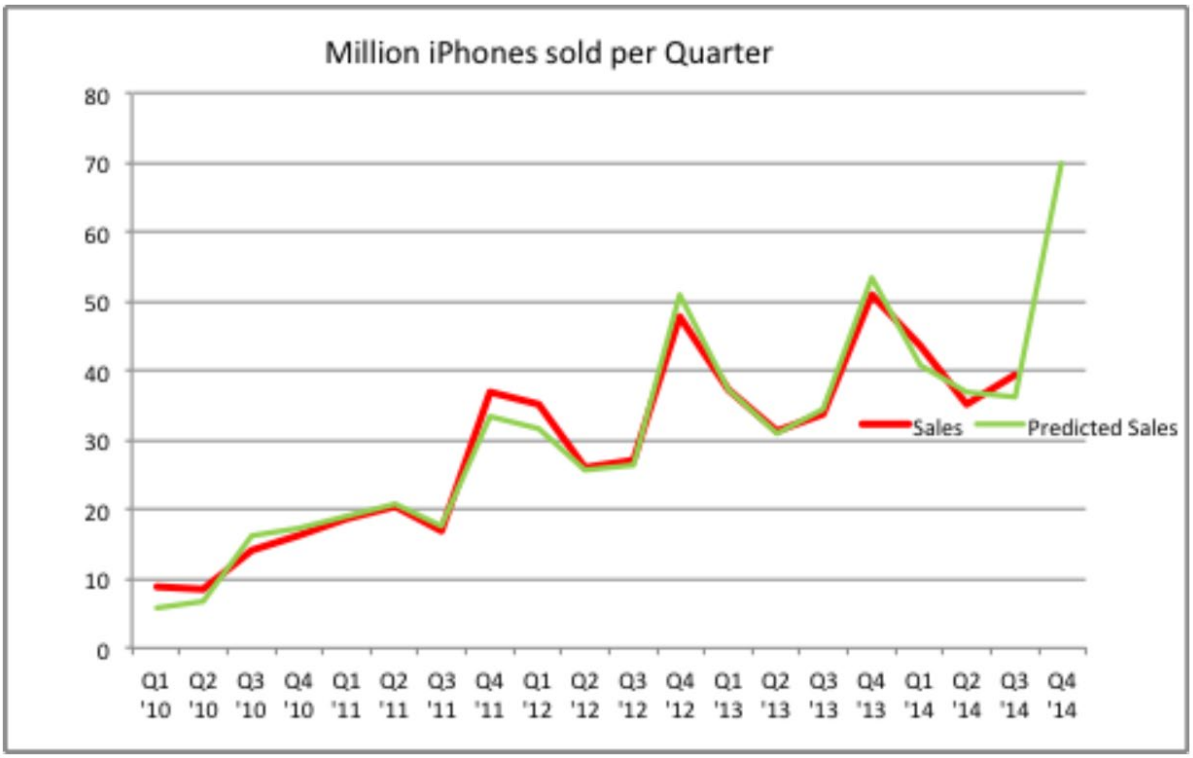


Figure 7: Q4.14 prediction of iPhone sales.

The AIDA and HOE customer journey models were used to explain the causal relationship between Twitter data to iPhone sales. Although these models are quite simple, this paper provides the first logical explanation of why social media data can predict sales.

Paper II: Predictive testing for the future was presented at several conferences in 2015. For the ICCSS 2015 Conference, a prediction of SEK 44 billion H&M sales for Q4.14 (H&M Q4 is September to November) was made by the ICCSS deadline of 7 December 2014.

The prediction of SEK 44 billion for Q4.14 was 3% over the actual sales of SEK 42.64 billion, for a prediction accuracy of 97%. The H&M Q4.14 sales on SEK 42.64 billion was published on 27 January 2015 at https://hmgroup.com/content/dam/hmgroup/groupsite/documents/en/cision/2015/01/1460341_en.pdf (retrieved 28 September 2020).

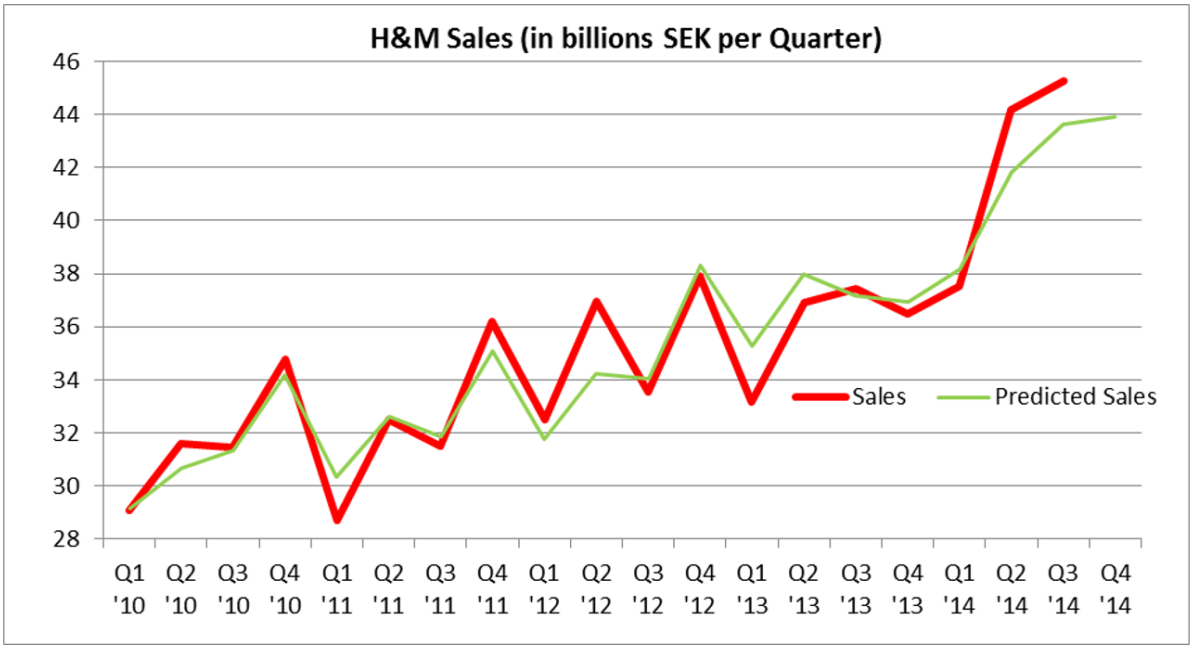


Figure 8: Q4.14 prediction of H&M sales from Facebook likes.

Paper III: Predictive testing was done on out-of-sample test data for July–September 2016. Mainly due to the relatively small dataset from small beer brand Mikkeller, the predictive power from Google data to Mikkeller beer sales was

very weak. There was an indication that Google searches for 'Mikkeller' lagged 5 months had some links to sales, but their predictive power was still weak.

A second version of this paper was presented at the same conference (Symposium i Anvendt Statistik) in January 2018 (la Cour et al. 2018), one year after the first version. In the second version, the Google data were replaced by Facebook data from all Mikkeller Facebook pages. The Facebook data also failed as predictors of Mikkeller sales, which confirms that the Mikkeller brand is too small to have enough social data to predict its sales. Further, paper II showed that Facebook data can predict sales if the brand and social data are large enough. Other predictive sales models have also succeeded by using Google searches, but for larger volumes products for which social data were large enough to predict sales (e.g. Choi and Varian 2012).

Paper IV: This is a review article of the datasets for 38 predictive models using social data that presents generalized predictive models for social data.

Paper V: Predictive testing was done on out-of-sample test data for the last half year of the 5-year dataset from 2015 to 2020. The predictive power for Google searches AAPL and AMZN for Apple stock volatility was relatively good, with a MAPE below 40, when forecasting up to February 2020. After February 2020, the predictive power of the model was affected by the effects of the COVID-19 pandemic on the financial markets during March–April 2020.

In conclusion, both Smueli et al (2011) and CRISP-DM, Wirth et al (2000), process modelling steps share the same iterative nature and emphasize the importance of continuous improvement and adaptation.

6. iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models

6.1 Choice of datasets

The iPhone dataset was chosen in paper I, because it was one of the most mentioned brands on Twitter. The iPhone brand is also unique and clean, compared to fx Samsung Galaxy and Google Pixel, which creates more noise. This is due to overlaps with too many Samsung products, and with Google Pixel there are overlaps with the Pixel word.
The dataset collected in paper I contained more than 500 million counts of tweets containing “iPhone”. This created the strongest predictive sales model in this PhD, due to the size of iPhone Twitter mentions, and the clean word nature of iPhone.

The data collection of more than 500 million counts of tweets containing “iPhone” was very time consuming, and there was no time or scope for collecting Twitter data on more brands in paper I. Applying the iPhone model on more brands, would of course have validated the model more, for claims about predictive power of Twitter data for sales.

Instead it was chosen to test Facebook and Google search data after the iPhone model, so the difference between Twitter, Facebook and Google searches could be analyzed.
This led to the choice of the H&M Facebook dataset and Mikkeller Google dataset.

These were the reasonings, for the choices of selecting iPhone Twitter and H&M Facebook datasets in paper I-II.
The Mikkeller dataset was chosen, because it was the only case where it was

possible to negotiate sales data on a monthly level, in contrast to the quarterly sales data for iPhone and H&M.
The Apple stock price volatility Google search dataset was chosen, because it was a chance to model on daily data for a five year period.

6.2 Limitations of linear regression models

Simple regression models, such as multiple linear regression models in paper I & II, can be insufficient for accurate prediction of sales based on social media data due to several reasons:

Non-linearity: Social media and web search data often exhibit non-linear relationships with sales, which cannot be captured by simple linear regression models. Eg. Chen et al (2018).

High dimensionality: Social media and web search data consist of numerous variables, such as text, images, timestamps, and user interactions, which can lead to high dimensionality. Simple regression models struggle to handle high-dimensional data and may result in overfitting. Eg. Blei et al (2003).

Feature interactions: Complex interactions between different variables in social media and web search data can be difficult to capture using simple regression models. Machine learning models, such as decision trees and ensemble methods, can effectively model such interactions. Eg. Breiman, L. (2001).

Temporal dependencies: Social media and web search data often exhibit temporal patterns, which can be essential for accurate sales prediction. Time series

models and recurrent neural networks are better suited for capturing these temporal dependencies. Eg. Hochreiter et al (1997).

Text analysis: Simple regression models cannot directly process text data, while machine learning models like topic models and word embeddings can extract useful information from textual content. Eg. Mikolov et al (2013).

The models chosen in paper I-II were multiple linear regression models, and I learned non-linear statistical and machine learning after these two papers, which changed the model choices in following papers.

6.3 Changing from linear models in paper I & II, to a non-linear model in paper III

These are some of the main reasons for changing from multiple regression to Unobserved Components Model (UCM) in paper III (predicting Mikkeller sales with Google searches), as the UCM model can deal with non-linear data relations, complex feature interactions and temporal dependencies. Eg. Harvey, A. C. (1989) and Durbin & Koopman (2012).

6.4 Changing from statistical models in paper I, II & III, to machine learning models in paper V

To overcome these limitations, machine learning models should be also considered for prediction of sales and stock price volatility based on social media data and web search data.
This was the main reason for choosing the Lasso regression (Least Absolute

Shrinkage and Selection Operator) model in paper V, as Lasso regression is a machine learning model. Lasso regression is a machine learning model because it involves learning from data to make predictions or inferences. It was first introduced by Robert Tibshirani in 1996 as a modification to traditional linear regression models, incorporating L1 regularization to improve the interpretability and generalization of the model (Tibshirani, 1996).

Lasso regression is a machine learning algorithm, which means it uses a dataset with input-output pairs to learn the that minimizes the mean squared error between the predicted and actual output values while also minimizing the sum of the absolute values of the coefficients (James, Witten, Hastie, & Tibshirani, 2013). This process of learning from data and optimizing the coefficients makes Lasso regression a machine learning model.

Lasso regression is particularly useful in scenarios where there are a large number of input variables and potential multicollinearity, as it performs both variable selection and regularization simultaneously. This helps prevent overfitting and improves the interpretability of the model by selecting only a subset of the most important features (Tibshirani, 1996).

In summary, Lasso regression is a machine learning model because it learns from data to make predictions, using regularization to improve generalization and interpretability.

Lasso regression can also deal with non-linear input data by introducing regularization to the model, specifically L1 regularization. Although Lasso itself is a linear model, it can be adapted to handle non-linear input data through various techniques, such as feature transformation, basis function expansion, or kernel methods.

a. Feature transformation: One common approach to handling non-linear input data is to apply non-linear transformations to the input features, creating new features that can better capture non-linear relationships in the data. For example,

polynomial features can be generated by squaring or cubing the original features, which can help Lasso capture non-linear patterns in the data (Friedman, Hastie, & Tibshirani, 2010).

b. Basis function expansion: Another approach is to use basis function expansion, such as splines or radial basis functions, to transform the input features into a higher-dimensional space. By applying L1 regularization to the expanded feature set, Lasso can still perform variable selection and model complex relationships in the data (Ravikumar, Lafferty, Liu, & Wasserman, 2009).

c. Kernel methods: Lasso can be combined with kernel methods, such as support vector machines, to learn non-linear relationships in the data.

In this method, a kernel function aids in translating the input data into a space with a greater number of dimensions, where linear relationships can be more easily modeled. Lasso can then be applied in this transformed space to perform variable selection and regularization (Zou, Hastie, & Tibshirani, 2006).

6.5 Comparing two machine learning models in paper V

In paper V, Lasso regression was compared with Oxmetrics Autometrics, see Doornik et al (2009).

Oxmetrics Autometrics is also a machine learning model.

OxMetrics Autometrics is a module developed by Sir David F. Hendry and Jurgen A. Doornik, which is part of the OxMetrics software package. OxMetrics is an econometric and statistical software that provides a wide range of tools for data manipulation, statistical analysis, and econometric modeling. The Autometrics module is designed for automatic model selection in econometric models, primarily for time-series data.

The Autometrics algorithm can be considered a machine learning model because it employs an iterative process to automatically search, select, and evaluate potential

explanatory variables for inclusion in the final econometric model. The methodology uses a general-to-specific approach, starting with a general model that includes all possible candidate variables, and iteratively reducing the model by excluding insignificant variables. The process continues until the most parsimonious and statistically significant model is selected.

The machine learning aspect of Autometrics lies in its ability to "learn" the best combination of variables for the given data and the specified model, based on reducing a measure of information like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to the smallest possible value. The procedure is data-driven, as it adapts to the data provided and the results improve as more data is incorporated.

Eg. Hendry et al (2005), Doornik et al (2009) and Hendry et al (2014).

These publications discuss the development, methodology, and applications of Autometrics in econometric modeling, highlighting its features as a machine learning model.

So two machine learning models, Lasso and Autometrics, were compared in paper V for the Apple dataset., predicting Apple stock price volatility with Google searches.

6.6 iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models

For the papers I, II, III and V, 40 new models were tested with Python's LazyPredict library, eg. Shah, S. (2020).

This was done to facilitate a discussion and analysis of the difference between statistical and machine learning models for the papers and datasets in this PhD.

Python's LazyPredict is a library that simplifies the process of comparing 40 statistical and machine learning models. Python's LazyPredict test all the models from scikit-learn (Pedregosa et al (2011) and Buitinck et al (2013).)

LazyPredict does this by automatically fitting and evaluating multiple models on a given dataset, providing a convenient way for data scientists and analysts to identify the best model for their problem. LazyPredict can be a good method for comparing these models for several reasons:

a. Saves time and effort: LazyPredict automates the process of fitting and evaluating multiple models, which saves time and effort that would otherwise be spent on writing code for each model individually (Varun, 2020).

b. Provides a comprehensive comparison: By quickly evaluating a wide range of models, LazyPredict helps users identify the best model for their specific use case. This comprehensive comparison is valuable because it ensures that the chosen model is well-suited to the problem at hand (Waskom, 2021).

c. Facilitates informed decision-making: LazyPredict generates a summary of key performance metrics (such as accuracy, precision, recall, F1 score, and others) for each model, enabling users to make data-driven decisions when selecting the most appropriate model for their problem (Sharma, 2020).

d. Encourages experimentation: LazyPredict encourages users to explore different models and understand their performance characteristics. This experimentation can lead to the discovery of new, more effective models or even to the development of hybrid models that combine the strengths of multiple algorithms (Singh, 2021).

In summary, LazyPredict was chosen for model testing, as it is a useful tool for comparing statistical and machine learning models. Because it automates the process, saves time, provides comprehensive comparisons, facilitates data-driven decision-making, and encourages experimentation.

A weakness of LazyPredict is the standard version of all the 40 models. Working with some of the best performing models from LazyPredict and fine tuning them, is of course a natural next step in accordance with the CRISP-DM process.

LazyPredict uses the scikit-learn formula for the adjusted R-square, which creates some R-squares less than minus 1 in the LazyPredict results for the iPhone, H&M, Mikkeller and Apple datasets in the following sections. This happens when the denominator in the below scikit-learn R-square formula get very small, then the fraction can result in a large negative number.

If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value for total n samples, the estimated R^2 is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\epsilon_i^2$.

Note that `r2_score` calculates unadjusted R^2 without correcting for bias in sample variance of y.

Figure 9, R-square formula from Scikit Learn

Source: *https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score*

Retrieved 16. May 2023.

6.7, LazyPredict 40 models results for paper 1, predicting iPhone sales with Twitter data

The Python LazyPredict coding for paper 1 can be found inAppendix 1, python

LazyPredict code for paper I, predicting iPhone sales with Twitter data, page 172. LazyPredict runs the iPhone dataset on a 80/20 train/test split, and displays the Adjusted R-square, R-square, RMSE and Time Taken for the test part of the dataset (table 10 below).

I will only analyze the best performing models of the 40 models tested, which means I will be comparing the top10 best models in table 10 against the multiple regression used in paper I. Before analyzing on the top10, I need to explain these top10 models shortly.

6.7.1 Bagging and boosting models

Bagging and boosting represent methods of ensemble learning, which are employed to enhance the efficiency of machine learning models, such as decision tree regressors, by combining multiple base learners to form a more accurate and robust model. Here is a brief explanation of each technique applied to decision tree regressors.

Bagging is an ensemble method that combines multiple decision tree regressors trained on different subsets of the training data. These subsets are generated by randomly sampling with replacement from the original dataset (bootstrap sampling). The final prediction is derived by calculating the mean of the individual predictions of individual decision tree regressors. Bagging reduces the variance of the model, making it less prone to overfitting. Random Forest is one of the most popular bagging versions (Breiman, 1996).

Boosting is another ensemble method that aims to improve the performance of weak learners, such as decision trees, by training multiple models sequentially. In each iteration, a new decision tree regressor is trained to correct the errors made by the previous model. Gradient Boosting is a machine learning technique used for supervised learning tasks, such as classification and regression. It involves

creating an ensemble of decision trees, where each tree is built to correct the errors of the previous tree. The technique was first introduced by Friedman in 2001 in his paper "Greedy Function Approximation: A Gradient Boosting Machine", Friedman et al (2001). Since then, it has become one of the most popular machine learning algorithms due to its high accuracy and flexibility.

6.7.2 Analysis

It can be seen in the results (table 10 below), that GradientBoosting has RMSE on 3.34, RandomForest has RMSE on 4.16, ADA Boost has RMSE on 4.9, and the LinearRegression used in paper 1, has RMSE on 9.12.

With an average of 37.4 million sold iPhones per quarter in the iPhone dataset, it means Gradient Boosting can predict out-of-sample with an average error of approximate 3.34 million iPhones per quarter, which is approximate 9% average prediction error. MAE on 2.46 million iPhones per quarter for Gradient Boosting is indicating approximate 6.6% average prediction error.

Random Forest would be approximate 11% average prediction error, and ADA Boost would be approximate 13% average prediction error, measured on RMSE. Measured on MAE, these two models would be approximate 9-12% average prediction error.

The LinearRegression predicts out-of-sample with an average error of approximate 9.12 million iPhones per quarter, which is approximate 24% average prediction error.
It is also worth noticing the R-square on 97% for GradientBoosting, 95% for Random Forest, 93% for ADA Boost and 77% for the LinearRegression used in paper 1. (table 10 below)

This indicates that boosting and bagging versions of the Decision Tree regression are working much better for predicting iPhone sales with Twitter data, compared to the Linear Regression used in paper 1.

Similar findings are done by Bifet et al (2010). "Mining Adaptive Micro-Clusters from Data Streams using Ensemble Methods"
This paper discusses the advantages of ensemble methods such as boosting and bagging for working with streaming data, which could be relevant when using Twitter data for predictions.
Similar findings are also done by Géron, A. (2019). "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" (2nd ed.). O'Reilly Media.
This book provides a comprehensive comparison of various machine learning models, including decision tree-based ensemble models and linear regression models, in terms of their predictive performance.
In general, boosting and bagging versions of the Decision Tree regression are very suitable for small datasets, and also performs well on non-linear data relations, which are my main explanations for the better performance of these models. This is also well-documented in the literature.
Eg. Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (2nd ed.). Springer. This widely-cited textbook covers various statistical learning methods, including ensemble methods based on decision trees, and provides evidence for their effectiveness in handling small datasets and non-linear relationships.
Refer also to Breiman, L. (1996). "Bagging Predictors". Machine Learning, 24(2). In this foundational paper, the author introduces the concept of bagging and demonstrates its ability to improve the performance of decision trees on small datasets.

Refer also to Freund, Y., & Schapire, R. E. (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". Journal of Computer and System Sciences, 55(1).

This paper introduces the AdaBoost algorithm, an example of a boosting ensemble method, and shows its effectiveness in improving the performance of decision trees in various settings, including those with non-linear relationships.

Below are the LazyPredict results for paper I, predicting iPhone sales with Twitter data.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken	mean_absolute_error
GradientBoostingRegressor	0.95	0.97	3.34	0.05	2.46
BaggingRegressor	0.94	0.96	3.67	0.03	2.99
DecisionTreeRegressor	0.94	0.96	3.81	0.01	3.04
RandomForestRegressor	0.93	0.95	4.16	0.14	3.45
ExtraTreesRegressor	0.92	0.95	4.43	0.14	3.78
AdaBoostRegressor	0.90	0.93	4.97	0.07	4.41
XGBRegressor	0.86	0.91	5.82	0.12	4.33
ExtraTreeRegressor	0.80	0.87	6.97	0.02	4.64
KNeighborsRegressor	0.79	0.86	7.05	0.03	5.23
PassiveAggressiveRegressor	0.73	0.82	8.01	0.01	6.00
HuberRegressor	0.72	0.82	8.16	0.04	6.20
Lars	0.66	0.77	9.12	0.02	7.31
TransformedTargetRegressor	0.66	0.77	9.12	0.01	7.31
LinearRegression	0.66	0.77	9.12	0.01	7.31
RidgeCV	0.65	0.77	9.14	0.01	7.39
Ridge	0.65	0.77	9.14	0.01	7.39
SGDRegressor	0.65	0.77	9.14	0.01	7.32
BayesianRidge	0.65	0.77	9.17	0.02	7.47
LassoCV	0.65	0.77	9.20	0.10	7.64
LassoLarsIC	0.65	0.77	9.20	0.01	7.59

LarsCV	0.65	0.77	9.21	0.03	7.67
LassoLarsCV	0.65	0.77	9.21	0.02	7.67
OrthogonalMatchingPursuitCV	0.65	0.77	9.21	0.02	7.51
Lasso	0.64	0.76	9.27	0.01	7.88
ElasticNetCV	0.64	0.76	9.33	0.07	7.86
PoissonRegressor	0.59	0.73	9.91	0.01	7.18
OrthogonalMatchingPursuit	0.58	0.72	10.05	0.01	7.81
RANSACRegressor	0.54	0.70	10.51	0.03	7.31
ElasticNet	0.51	0.67	10.93	0.01	9.02
GammaRegressor	0.47	0.65	11.28	0.02	9.15
LassoLars	0.39	0.59	12.13	0.01	9.89
TweedieRegressor	0.37	0.58	12.39	0.01	9.98
LinearSVR	0.16	0.44	14.23	0.01	10.97
HistGradientBoostingRegressor	0.01	0.34	15.51	0.14	13.90
NuSVR	-0.06	0.29	16.02	0.01	12.94
SVR	-0.06	0.29	16.04	0.01	12.51
LGBMRegressor	-0.24	0.17	17.31	0.05	15.90
DummyRegressor	-0.54	-0.03	19.30	0.01	15.70
MLPRegressor	-2.44	-1.30	28.85	0.14	23.42
GaussianProcessRegressor	-3.68	-2.12	33.64	0.02	19.06
KernelRidge	-3.84	-2.23	34.23	0.02	33.32

Table 10, LazyPredict results for paper I, predicting iPhone sales with Twitter data

6.8, LazyPredict 40 models results for paper II, predicting H&M sales with Facebook data

The Python LazyPredict coding for paper II can be found in Appendix 2, python LazyPredict code for paper II, predicting H&M sales with Facebook data, page 174.

LazyPredict runs the H&M dataset on a 65/35 train/test split, and displays the Adjusted R-square, R-square, RMSE, Time Taken and MAE for the test part of the dataset (table 11 below).

I will only analyze the best performing models of the 40 models tested, which means I will be comparing the top6 best models in table 11 against the multiple regression used in paper II. Before analyzing on the top6, I need to explain these top6 models shortly.

6.8.1 Regularized and generalized linear models (GLM)

Tweedie regression is a type of generalized linear model (GLM) that is particularly useful for modeling non-negative, continuous, and discrete data, especially when the data exhibits both zero and non-zero values. It belongs to the family of exponential dispersion models (EDMs) and is named after the Scottish statistician Maurice Tweedie (Tweedie, 1984). The Tweedie distribution is characterized by a power parameter (p) that determines the distribution's properties. Depending on the value of p, the Tweedie distribution can represent various distributions, such as normal (p=0), Poisson (p=1), gamma (p=2), and

inverse Gaussian (p=3) (Jorgensen, 1987).

Elastic Net is a regularization technique used in linear regression models to address multicollinearity and overfitting. It combines two popular regularization methods: L1-norm (Lasso) and L2-norm (Ridge) regularization. This hybrid approach effectively penalizes large coefficients and enforces sparsity in the model, improving generalization performance and interpretability. Elastic Net was first introduced by Hui Zou and Trevor Hastie in their 2005 paper titled "Regularization and variable selection via the elastic net" (Zou & Hastie, 2005).

RANSAC (RANdom SAmple Consensus) Regressor is a robust regression algorithm that is particularly effective in dealing with noisy, outlier-prone datasets. It is an iterative method for estimating a mathematical model from a dataset that contains a significant number of outliers (Fischler and Bolles, 1981). RANSAC is a popular method in computer vision and robotics applications, such as 3D reconstruction and image stitching.

GammaRegressor is a regression model used for predicting continuous target variables when the underlying distribution is assumed to follow the Gamma distribution. It is particularly useful when the data contains a skewed distribution, with many smaller values and fewer larger values (Gamma distribution). The GammaRegressor is based on the Generalized Linear Model (GLM) framework with a Gamma distribution as the response (McCullagh & Nelder, 1989). The model is estimated using a maximum likelihood approach.

Huber Regressor is a linear model for regression that is robust to outliers, combining the best properties of both linear regression and robust regression methods. It was first introduced by Peter J. Huber in 1964 in his seminal paper

titled "Robust Estimation of a Location Parameter" (Huber, 1964).

Ridge regression, also known as L2 regularization, is a technique used in linear regression to address the problem of multicollinearity in the independent variables. It was first introduced by Arthur E. Hoerl and Robert W. Kennard in their 1970 paper titled "Ridge Regression: Biased Estimation for Nonorthogonal Problems", (Hoerl, A. E., & Kennard, R. W. 1970). Ridge regression has been widely adopted in various applications such as finance, medical research, and engineering, where multicollinearity is a common issue.

6.8.2 Analysis

It can be seen in the results (table 11 below), that ElasticNet and TweedieRegressor has RMSE on approximate 1,5 billion SEK per quarter. RANSARegressor and GammaRegressor has RMSE on approximate 1,6 billion SEK per quarter. HuberRegressor and Ridge has RMSE on approximate 1,7 billion SEK per quarter.
The LinearRegression used in paper II, has RMSE on 1,8 billion SEK per quarter. It should be noted here, the linear regression in paper II, had also RMSE on 1,8 billion SEK per quarter, see paper II, section II.3. Methodology and analytical findings, figure 5, page 206.

With an average of 34,56 billion SEK H&M sales per quarter in the H&M dataset, it means ElasticNet and TweedieRegressor can potentially predict out-of-sample with an average error of approximate 1,5 billion SEK per quarter, which is approximate 4% average prediction error. MAE for these two models are 1.2

billion SEK per quarter, indicating approximate 3.5% average prediction error.

RANSARegressor, GammaRegressor, HuberRegressor and Ridge can potentially predict out-of-sample with an average error of approximate 1,7-1,8 billion SEK per quarter, which is approximate 5% average prediction error. MAE for these models are 1.4 - 1.5 billion SEK per quarter, indicating approximate 4 - 4.5% average prediction error.

The Linear Regression used in paper II can potentially predict out-of-sample with an average error of approximate 1,8 billion SEK per quarter, which is approximate 5% average prediction error. In practice this would not be possible because the model Linear Regression is weak with a R-square on 28%. It should be noted that, the r-square for this model is 83% in paper II, because timelag of Facebook data and also seasonal weights were applied. See paper II, section II.3. Methodology and analytical findings, figure 5, page 206.

It is also worth noticing the R-square on 54% for TweedieRegressor, 53% for ElasticNet, 45% for RANSACRegressor, 42% for GammaRegressor, 38% for HuberRegressor, 37% for Ridge and 28% for the Linear Regression (table 11 below)
Together with RMSE numbers above, this is indicating that the TweedieRegressor, ElasticNet and RANSACregressor works better for predicting H&M sales with Facebook data, compared to the Linear Regression used in paper II.
Compared to the overweight of bagging and boosting models for the top10 of iPhone dataset in table 10, we see here for the H&M dataset an overweight of regularized and general linear models. There is more non-linearity in Twitter data

relations to sales, compared to Facebook data relations to sales, and that is my explanation for mainly linear models in the top10 for H&M.
The Twitter dataset captures all the tweets containing “iPhone”, where the H&M dataset only captures Facebook likes on the H&M Facebook page. So the iPhone Twitter dataset is measuring product attention much broader than the H&M dataset. Twitter data are also less filtered than Facebook data, which is explained in the chapter 7. Social Filtering Model.
This means Twitter data captures more raw product attention than Facebook data.

It is also worth noticing, that r-squares were much stronger in table 10, LazyPredict results for the iPhone dataset (paper I). The 7 best models for the iPhone dataset in table 10, had r-square >90%. This is again indicating, that Twitter data are better for sales predictions, compared to Facebook data.
Below are the LazyPredict results for paper II, predicting H&M sales with Facebook data.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken	mean_absolute_error
TweedieRegressor	0.26	0.54	1471.12	0.01	1194.01
ElasticNet	0.24	0.53	1488.24	0.01	1180.18
RANSACRegressor	0.12	0.45	1599.73	0.01	1397.03
GammaRegressor	0.08	0.42	1640.95	0.01	1354.36
HuberRegressor	0.01	0.38	1698.39	0.04	1490.72
Ridge	-0.00	0.37	1707.71	0.01	1535.15
BayesianRidge	-0.02	0.36	1727.25	0.01	1556.99
ElasticNetCV	-0.07	0.33	1769.93	0.05	1644.79
RidgeCV	-0.13	0.29	1818.55	0.01	1645.47
SGDRegressor	-0.15	0.28	1827.46	0.01	1652.99
Lars	-0.15	0.28	1834.78	0.02	1659.37
LinearRegression	-0.15	0.28	1834.78	0.01	1659.37
TransformedTargetRegressor	-0.15	0.28	1834.78	0.01	1659.37
Lasso	-0.16	0.28	1835.01	0.01	1659.67
LassoLars	-0.16	0.28	1835.44	0.01	1660.35
LassoCV	-0.18	0.27	1851.24	0.05	1633.34
LarsCV	-0.20	0.25	1866.93	0.01	1680.68
LassoLarsCV	-0.20	0.25	1866.93	0.01	1680.68
LassoLarsIC	-0.23	0.23	1892.62	0.01	1704.70
OrthogonalMatchingPursuitCV	-0.26	0.21	1919.46	0.01	1734.84

PoissonRegressor	-0.35	0.16	1983.66	0.01	1819.00
LGBMRegressor	-0.65	-0.03	2196.30	0.03	2002.78
DummyRegressor	-0.65	-0.03	2196.30	0.01	2002.78
HistGradientBoostingRegressor	-0.65	-0.03	2196.30	0.05	2002.78
NuSVR	-1.18	-0.36	2518.70	0.01	2113.86
XGBRegressor	-1.73	-0.71	2821.61	0.06	2420.64
ExtraTreesRegressor	-1.91	-0.82	2910.45	0.08	2602.64
OrthogonalMatchingPursuit	-2.42	-1.13	3155.48	0.01	2636.75
SVR	-2.42	-1.14	3158.85	0.01	2506.41
KNeighborsRegressor	-2.65	-1.28	3262.26	0.02	2866.11
RandomForestRegressor	-3.34	-1.71	3556.43	0.09	3042.43
BaggingRegressor	-4.00	-2.13	3817.83	0.02	3323.37
AdaBoostRegressor	-4.21	-2.25	3895.77	0.06	3350.30
GradientBoostingRegressor	-4.28	-2.30	3922.64	0.03	3297.00
ExtraTreeRegressor	-6.23	-3.52	4592.22	0.01	4062.44
DecisionTreeRegressor	-6.94	-3.96	4809.91	0.01	4076.00
GaussianProcessRegressor	-27.64	-16.90	9137.98	0.01	7438.97
PassiveAggressiveRegressor	-135.23	-84.14	19927.82	0.01	19810.44
KernelRidge	-415.77	-259.48	34856.04	0.02	34816.55
LinearSVR	-415.93	-259.58	34862.40	0.01	34795.44
MLPRegressor	-416.07	-259.67	34868.44	0.07	34801.53

Table 11, LazyPredict results for paper II, predicting H&M sales with Facebook data

6.9, LazyPredict 40 models results for paper III, predicting Mikkeller sales with Google search data

The Python LazyPredict coding for paper III can be found in Appendix 3, python LazyPredict code for paper III, predicting Mikkeller beer sales with Google searches, page 180.
LazyPredict runs the Mikkeller dataset on a 80/20 train/test split, and displays the Adjusted R-square, R-square, RMSE, Time Taken and MAE for the test part of the dataset (table 12 below).
I will only analyze the best performing models of the 40 models tested, which means I will be comparing the top8 best models in table 12 against the UCM time series model used in paper III. Before analyzing on the top8, I need to explain these top8 models shortly.

6.9.1 Adaboost, regularized and generalized linear models (GLM)

The Adaboost regressor is explained in chapter 6.7.1 Bagging and boosting, in the iPhone analysis above.
The HuberRegressor is is explained in chapter 6.8.1 Regularized and generalized linear models (GLM), in the H&M analysis above.

The LARS (Least Angle Regression and Selection) model is a statistical method for linear regression that is particularly effective when dealing with high-dimensional data. The algorithm's main advantage is its ability to select relevant features for prediction while maintaining a low computational cost compared to other feature selection methods.
The LARS algorithm is closely related to other regularization methods like

LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge Regression. LARS can be modified to implement LASSO by adding a constraint on the L1-norm of the coefficients. This modification leads to a more sparse solution, as some coefficients are forced to be exactly zero. (Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. 2004).

TransformedTargetRegressor is a machine learning technique that applies a transformation to the target variable before fitting the regression model and inverts the transformation when predicting the output. This approach is particularly useful when the relationship between the input and output variables is non-linear, or when the target variable has a skewed distribution. By transforming the target variable, the model can better capture the underlying patterns in the data, potentially leading to improved performance (Pedregosa et al 2011).

Stochastic Gradient Descent (SGD) Regressor is a linear model that uses Stochastic Gradient Descent for optimization. It is a widely used algorithm in machine learning, particularly in large-scale and sparse data problems. The model can be used for regression tasks, where the goal is to predict a continuous target variable based on input features.

SGD Regressor works by iteratively updating the model's weights using a random subset of the data, called mini-batches, to minimize the objective function (typically the Mean Squared Error). This approach offers faster convergence compared to traditional Gradient Descent, which computes the gradient using the entire dataset in each iteration.
(Bottou, 2010; Zhang, 2004).

The Poisson regressor is a statistical model used for predicting count outcomes by

modeling the relationship between a set of explanatory variables and a non-negative integer response variable. It belongs to the family of generalized linear models (GLMs) and assumes that the response variable follows a Poisson distribution. The Poisson distribution is commonly used to model rare events or occurrences with a constant rate of occurrence in a fixed time or space interval **(**McCullagh 1989**;** Cameron 2013).

The Ridge model is explained in chapter 6.8.1 Regularized and generalized linear models (GLM), in the H&M analysis above.

6.9.2 Analysis

It can be seen in the results (table 12 below), that AdaBoostRegressor has RMSE on 11,58 and R-square on 44%.
It is very interesting to see the Adaboost as the best performing model, as the boosting models were also very present in the top10 best perfoming models of the iPhone dataset, see Table 10, LazyPredict results for paper I, predicting iPhone sales with Twitter data.
My explanation for boosting models in top10 best performing models for both iPhone and Mikkeller dataset, is that Twitter and Google search data relations to sales are more non-linear, compared to Facebook data relations to sales.
Twitter and Google search data are also less filtered than Facebook data, which is explained in the chapter 7. Social Filtering Model.
This means Twitter and Google search data captures more raw product attention than Facebook data.
Like the H&M dataset, we also find the Huber and Ridge models in the top10 of best performing models.

HuberRegressor, Lars, TransformedTargetRegressor, SGDRegressor, PoissonRegressor, Ridge and Linear Regression have RMSE just around 12.
The Mikkeller sales were indexed from 20-100, because Mikkeller did not want to reveal their sales data.
With an average sales per month in the Mikkeller dataset on index 50, it means the AdaboostRegressor can potentially predict out-of-sample with an average error of approximate index 11,58, which is approximate 23% average prediction error.
MAE on 8.53 for the AdaboostRegressor is indicating approximate 17% average prediction error.
It should be noted, that AdaboostRegressor is relatively weak with a R-square on 44%.
This indicates that the AdaboostRegressor, and some regularized and general linear models from top10, could be a little better alternative to the UCM time series model used in paper III.

HuberRegressor, Lars, TransformedTargetRegressor, SGDRegressor, PoissonRegressor, Ridge and Linear Regression can potentially predict out-of-sample with an average error of approximate index 12 which is approximate 24% average prediction error. MAE on approximate 9.5 for these models are indicating approximate 19% average prediction error. It should be noted, that these models are relatively weak with a R-squares on 38%-39%.

The R-squares for the Mikkeller dataset in top10 are relative comparable with the R-squares for the H&M dataset in Table 11, LazyPredict results for paper II, predicting H&M sales with Facebook data.
Below are the LazyPredict results for paper III, predicting Mikkeller sales with Google search data.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken	mean_absolute_error
AdaBoostRegressor	-2.38	0.44	11.58	0.07	8.53
HuberRegressor	-2.64	0.39	12.02	0.02	9.17
Lars	-2.65	0.39	12.03	0.02	9.20
TransformedTargetRegressor	-2.65	0.39	12.03	0.01	9.20
LinearRegression	-2.65	0.39	12.03	0.01	9.20
SGDRegressor	-2.66	0.39	12.05	0.01	9.23
PoissonRegressor	-2.73	0.38	12.16	0.01	9.53
Ridge	-2.73	0.38	12.17	0.01	9.32
OrthogonalMatchingPursuitCV	-2.87	0.35	12.40	0.01	10.92
OrthogonalMatchingPursuit	-2.87	0.35	12.40	0.01	10.92
Lasso	-2.96	0.34	12.54	0.01	9.81
BayesianRidge	-3.15	0.31	12.84	0.01	9.71
RidgeCV	-3.28	0.29	13.03	0.01	9.79
LassoCV	-3.28	0.29	13.04	0.05	10.44
LassoLars	-3.30	0.28	13.07	0.01	10.68
PassiveAggressiveRegressor	-3.32	0.28	13.10	0.01	9.18
LarsCV	-3.35	0.28	13.13	0.01	10.54
LassoLarsCV	-3.35	0.28	13.13	0.01	10.54
ElasticNetCV	-3.43	0.26	13.26	0.06	10.09
LassoLarsIC	-3.46	0.26	13.30	0.01	10.85

ElasticNet	-3.46	0.26	13.30	0.01	10.13
GammaRegressor	-3.50	0.25	13.36	0.01	10.11
BaggingRegressor	-3.71	0.22	13.67	0.02	10.88
TweedieRegressor	-3.85	0.19	13.87	0.01	10.43
ExtraTreesRegressor	-4.27	0.12	14.47	0.09	12.05
KNeighborsRegressor	-4.55	0.07	14.85	0.02	11.65
NuSVR	-4.78	0.04	15.15	0.01	11.53
RandomForestRegressor	-4.91	0.01	15.32	0.09	12.52
XGBRegressor	-5.17	-0.03	15.65	0.06	14.28
GradientBoostingRegressor	-5.23	-0.04	15.73	0.03	12.93
DummyRegressor	-6.14	-0.19	16.83	0.01	13.34
HistGradientBoostingRegressor	-6.14	-0.19	16.83	0.04	13.34
LGBMRegressor	-6.14	-0.19	16.83	0.03	13.34
SVR	-6.14	-0.19	16.83	0.01	13.29
ExtraTreeRegressor	-8.42	-0.57	19.34	0.01	15.99
RANSACRegressor	-8.95	-0.66	19.88	0.02	13.40
GaussianProcessRegressor	-8.96	-0.66	19.88	0.01	14.65
DecisionTreeRegressor	-14.63	-1.61	24.91	0.01	18.00
LinearSVR	-14.94	-1.66	25.16	0.01	21.36
MLPRegressor	-35.70	-5.12	38.17	0.08	35.86
KernelRidge	-48.58	-7.26	44.37	0.02	43.43

Table 12, LazyPredict results for paper III, predicting Mikkeller sales with Google search data

6.10, LazyPredict 40 models results for paper V, predicting Apple stock volatility with Google search data

The Python LazyPredict coding for paper V can be found in Appendix 4, python LazyPredict code for paper V, predicting Apple stock price volatility with Google searches, page 186.

LazyPredict runs the Apple dataset on a 80/20 train/test split, and displays the Adjusted R-square, R-square, RMSE, Time Taken and Mean Absolute Error for

the test part of the dataset (table 13 below).
I will only analyze the best performing models of the 40 models tested, which means I will be comparing the top10 best models in table 13 against the Lasso and Autometrics model used in paper V. Before analyzing on the top10, I need to explain these top10 models shortly.

6.10.1 Boosting and bagging, OMP, kNN and Transformed Target models

The boosting and bagging models in the top10 of Table 13, LazyPredict results for paper V, predicting Apple stock price volatility with Google search data, are explained in chapter 6.7.1 Bagging and boosting, in the iPhone analysis above.

The k-Nearest Neighbors (k-NN) regressor is a non-parametric machine learning algorithm used for regression tasks. It was first introduced by Evelyn Fix and Joseph Hodges (Fix, E., & Hodges, J. L. 1951).
The k-NN regressor is a simple yet powerful algorithm based on the concept of similarity. Given a new input, it finds the k training examples that are most similar to the input and predicts the output based on the average of the outputs of these k nearest neighbors (Cover, T., & Hart, P. 1967)

Orthogonal Matching Pursuit CV (OMPCV) is a model selection method for Orthogonal Matching Pursuit (OMP) that uses cross-validation to determine the optimal number of non-zero coefficients (sparsity) for the model. OMP is a greedy algorithm used for sparse signal recovery or sparse approximation, which has applications in areas such as signal processing, compressed sensing, and machine learning. (Pati, Y. C., Rezaiifar, R., & Krishnaprasad, P. S. 1993).

TransformedTargetRegressor is explained in chapter 6.9.1 Adaboost, regularized and generalized linear models (GLM) in the Mikkeller analysis above.

6.10.2 Analysis

It can be seen in the results (table 13 below), that all top10 performing models in the LazyPredict has RMSE on 0.02. With an average stock price volatility in the Apple dataset on 3.12%, it is a relatively high RMSE.
In paper V, the RMSE is 0.034 for Autometrics out-of-sample, and RMSE is 0.01 when excluding the COVID pandemic period. See section V.7. Forecasting evaluation, page 314, table 4.
With RMSE on 0.02 out-of-sample for the top10 performing models in the LazyPredict results (table 13 below), including the COVID pandemic period, it can be concluded boosting and bagging models are out-performing the Autometrics model for the Apple stock price volatility Google search dataset.

The Mean Absolute Error is 0.01 for top25 of the best performing models, indicating these models could potentially predict with an average error of approximate 33%.

AdaBoostRegressor has r-square on 0.66, XBGRegressor has r-square on 0.64, GradientBoostingRegressor has r-square on 0.63, ExtraTreesRegressor has r-square on 0.62 and RandomForestRegressor has r-square on 0.58.
Like the LazyPredict models in chapter 6.7, LazyPredict 40 models results for paper 1, predicting iPhone sales with Twitter data, we see mainly boosting and bagging versions of the decision tree regressor in the top6 performing models of LazyPredict.
After top6, we see the KNeighborsRegressor with R-square on 53%, Orthogonal

Matching Pursuit and Lars with R-square 52%.

The r-squares for the top10 performing models in the LazyPredict table 13, are on the level of r-squares in the paper V with Lasso and Autometrics modelling.

This is indicating that boosting and bagging versions of the decision tree regressor, should also be tested for financial predictive modelling with Google searches, together with more classical models like fx. the used Lasso and Autometrics models used in paper V.
Below are the LazyPredict results for paper V, predicting Apple stock price volatility with Google search data.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken	mean_absolute_error
AdaBoostRegressor	0.62	0.66	0.02	0.14	0.01
XGBRegressor	0.60	0.64	0.02	0.21	0.01
GradientBoostingRegressor	0.59	0.63	0.02	0.14	0.01
ExtraTreesRegressor	0.58	0.62	0.02	0.38	0.01
RandomForestRegressor	0.54	0.58	0.02	0.53	0.01
DecisionTreeRegressor	0.50	0.55	0.02	0.03	0.01
KNeighborsRegressor	0.47	0.53	0.02	0.03	0.01
OrthogonalMatchingPursuitCV	0.47	0.52	0.02	0.05	0.01
Lars	0.47	0.52	0.02	0.04	0.01
TransformedTargetRegressor	0.47	0.52	0.02	0.03	0.01
LassoLarsIC	0.47	0.52	0.02	0.07	0.01
LarsCV	0.47	0.52	0.02	0.05	0.01
LassoLarsCV	0.47	0.52	0.02	0.09	0.01
LinearRegression	0.47	0.52	0.02	0.06	0.01
LassoCV	0.47	0.52	0.02	0.19	0.01
ElasticNetCV	0.47	0.52	0.02	0.22	0.01
Ridge	0.47	0.52	0.02	0.02	0.01
BayesianRidge	0.46	0.52	0.02	0.06	0.01
RidgeCV	0.46	0.51	0.02	0.02	0.01
SGDRegressor	0.45	0.51	0.02	0.03	0.01

BaggingRegressor	0.44	0.49	0.02	0.10	0.01
LGBMRegressor	0.43	0.49	0.02	0.09	0.01
OrthogonalMatchingPursuit	0.43	0.49	0.02	0.03	0.01
HistGradientBoostingRegressor	0.43	0.49	0.02	0.56	0.01
HuberRegressor	0.42	0.48	0.02	0.07	0.01
LinearSVR	0.41	0.46	0.02	0.10	0.01
ExtraTreeRegressor	0.41	0.46	0.02	0.06	0.02
GammaRegressor	0.35	0.41	0.02	0.03	0.01
TweedieRegressor	0.30	0.37	0.02	0.03	0.01
NuSVR	0.30	0.37	0.02	0.07	0.02
PoissonRegressor	-0.07	0.04	0.03	0.02	0.02
Lasso	-0.11	-0.00	0.03	0.03	0.02
DummyRegressor	-0.11	-0.00	0.03	0.02	0.02
ElasticNet	-0.11	-0.00	0.03	0.03	0.02
LassoLars	-0.11	-0.00	0.03	0.05	0.02
RANSACRegressor	-0.12	-0.01	0.03	0.22	0.02
PassiveAggressiveRegressor	-0.84	-0.66	0.04	0.03	0.03
KernelRidge	-0.91	-0.72	0.04	0.03	0.03
SVR	-3.06	-2.67	0.05	0.02	0.05
MLPRegressor	-8.98	-8.00	0.08	0.35	0.06
GaussianProcessRegressor	-24.49	-21.99	0.13	0.04	0.10

Table 13, LazyPredict results for paper V, predicting Apple stock price volatility with Google search data

6.11 Conclusions for the LazyPredict models on iPhone, H&M, Mikkeller and Apple datasets

As mentioned in beginning of this chapter, it is a weakness of LazyPredict, that it uses the standard version of all the 40 models. Working with some of the best performing models from LazyPredict and fine tuning them, is of course a natural next step in accordance with the CRISP-DM process. The relative poor performance of LazyPredict linear regression for the iPhone and H&M datasets, should also be seen in the light of much better performing linear regression models in papers I-II. This is of course due to the fact, that fine tuning the linear models in papers I-II included timelag experiments of the social media data, and also the use of seasonal weights.

It looks like boosting and bagging versions of the decision tree regressor, are working much better for predictive sales models using Twitter data, compared to linear models.
The boosting and bagging models are also better than other machine learning models in the LazyPredict results, but fine tuning could of course change that.
The AdaBoost model was also the best performing LazyPredict model for the Mikkeller Google search dataset, and boosting and bagging models were also the best top6 LazyPredict models for the Apple stock price volatility Google search dataset.

Google searches and Twitter data are classified as low to medium filtered social data in the following chapter 7. Social Filtering Model, and it looks like boosting and bagging versions of the decision tree regressor works good for these data.

Facebook data are classified as highly filtered social data in the following chapter 7. Social Filtering Model, and this could be one of the reasons for the lack of boosting and bagging models in the LazyPredict best models for the H&M Facebook dataset.

Another reason could be, the limitation of the H&M dataset only capturing Facebook likes on the H&M Facebook page. It means the other datasets with Twitter and Google search data, captures more broad product attention than the H&M Facebook dataset.

7. Social Filtering Model

This PhD thesis tests the predictive power of data from Twitter, Facebook, Google and YouTube searches for sales modelling in Papers I–III. The differences in predictive power are mainly due to the social filtering of data on the different platforms, that is, we show our identities using different social filters, depending on the social platform.

These differences in the nature of social data led to the development of the Social Filtering Model, which explains the social filters applied on social media, blogs, forums, and web search data. Paper IV includes a description of the social filtering differences between social media platforms and also between blogs, forums, and web searches.

The idea to this model came from the differences in the predictive power between social media platforms and web search data. For instance, there is a significant difference between the predictive power of Twitter data in Paper I and that of Facebook data in Paper II. Both Facebook and Google search data failed in Paper III to explain Mikkeller beer sales, because the datasets were too small. There are no academic citations for predictive modelling of beer sales with social data, but

master thesis' I supervised, proved that Google search data work for predictive modelling for Carlsberg beer sales. So both the size and source of data matter. These projects also showed that Google Trends search data work better for the predictive modelling of sales compared to Instagram and Facebook. Google Trends search data are available in indexed form for free, with no API restrictions, and the Social Filtering Model will also show the predictive advantages of these data. The practical predictive modelling supervision of several hundreds of student projects and my master thesis led to the use of Google Trends data in Paper V. I would have chosen Google Trends search data for modelling iPhone and H&M sales in Papers I and II, if I were to propose those models again.

In short, the difference in predictive power between the different data sources depends of how filtered the data are, which the Social Filtering Model explains and conceptualises. Specifically, when we observe the actions of others on social media, web searches, blogs, and forums, how can we determine the degree of transparency? These different data types show filtering differences that determine their potential for predictive modelling.

The degree of data filtering expresses how close we are to the functioning of the human brain, that is, what individuals actually mean. Web searches will always be closer to the truth, as we can search for anything we want. These data are only filtered if we web search next to other people and care about their opinion or worry about employers monitoring our web searches. This explains why web search data often have a higher predictive modelling potential compared to more filtered or polished data.

Twitter will always have a medium degree of data filtering, as many people care about their identities on Twitter and try to attract likes, retweets, and replies. Some tweets have a low degree of data filtering, that is, when people are expressing their

opinions and do not care about other people's reactions or opinions. Twitter also has the reputation of often being the first in sharing some news; that is, being first or early in reporting of news is important, while grammar and spelling errors become less important. However, the hunt for likes, retweets, and replies creates filtered and polished data, which is missing from web searches. Blogs and forums have the same medium filtering, as people are also hunting likes, views, and re-posts on these channels.

Instagram is a good example of highly filtered and polished data. Instagrammers often measure their success by the number of followers, likes, reposts, and comments. As such, many Instagram posts are intended to get as many reactions as possible. Role models and influencers on Instagram and TikTok with a high number of followers have established a new economic system, where it is possible to make a living from posting about products, causes, organisations, and brands. However, negative consequences of Instagram on young people's mental health have been reported since 2010, when Instagram appeared (Wells et al. 2021). TikTok and Facebook share the same highly filtered and polished data. People tend to display successes on these platforms, presenting a highly polished picture of their lives. Perceived failures in life, such as depression, mental problems, alcohol and drug abuse, stress, break downs, economic problems, divorces, and being fired, are not popular topics to display on these channels.

These differences in the nature of social data, measured as their degree of filtering, determine their potential use in predictive models and explain why Google Trends is often a relevant benchmark. To illustrate some of these differences, I introduce Figure 9: Social Filtering Model.

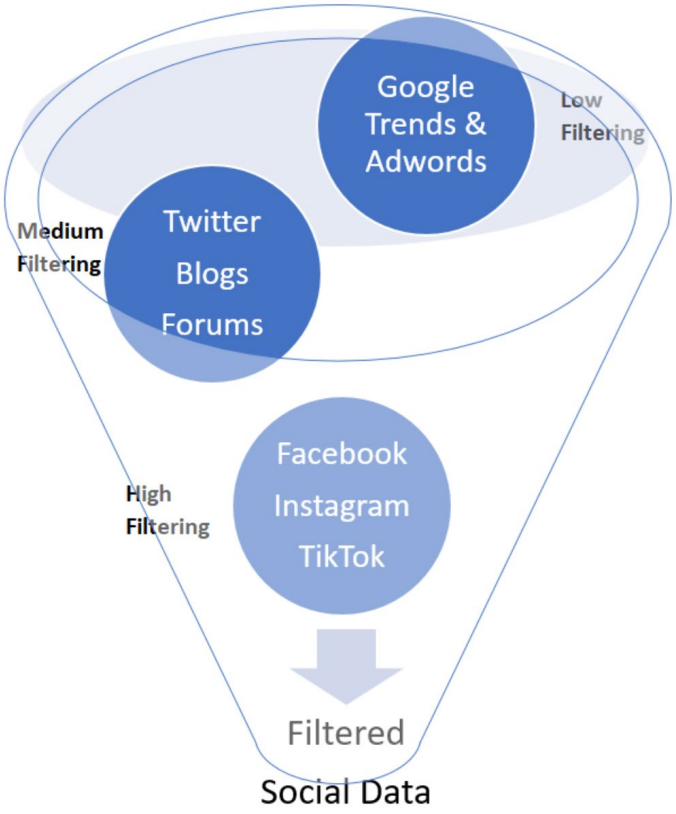


Figure 10: Social Filtering Model.

7.1 Dimensions of social media filtering

Social media data are filtered under several dimensions, with the main ones explained below:

a) The filters individuals apply to the content they publish;
b) The filter bubble algorithms social media and web search sites use to personalise content based on users' interest;
c) The filter social media and web search sites use for API access to their data.

7.1.a The filters individuals apply to the content they publish

During the research process, I have observed important filtering differences between social media platforms and web searches, but the literature proposes no theoretical or practical model for analysing these differences.
The filtered images of other peoples' lives on social media have been researched and documented for more than a decade, refer to Chou et al (2012), Fardouly et al (2015), Fardouly et al (2016), Haferkamp et al (2011) and Vogel (2014).

Still there is hitherto no model of the differences in filtering among different social media platforms in terms of either online norms of negative and positive emotions or social comparison. These two dimensions are covered below.

7.1.a.1 Online norms for negative and positive emotions on social media

The online norms for negative emotions on social media refer to the expectations and rules that govern how individuals express and respond to negative emotions on social platforms. These norms vary by platform and community, but they generally involve a balance between expressing oneself freely and being respectful

to others.

One norm common across many social media platforms is that individuals are expected to be respectful and considerate when expressing negative emotions. This includes avoiding hate speech or personal attacks, as well as being mindful of the tone and language used when expressing negative emotions. For instance, research conducted by the Pew Research Center demonstrated that the majority of social media users believe that it is never acceptable to use hate speech or racial slurs on social media (Perrin 2016).

Another norm that is often seen on social media is that individuals are expected to be open and honest when expressing negative emotions. This can involve sharing personal experiences and struggles, as well as being open to receiving support and feedback from others. A study conducted at the University of California, Berkeley found that individuals who had personal experience of mental health struggles on social media received more support and engagement from their peers than those who did not share such experiences (Ritter et al. 2016).

It is also important to note that online norms for negative emotions on social media can vary by platform and community. For example, a study from the University of Missouri found that the norms for expressing negative emotions on Instagram are different than those on Twitter, with Instagram users being more likely to express negative emotions in a more positive or humorous manner (Garcia and Mirra 2019).

Another study found that social media users tend to self-censor their negative emotions, and instead present a curated version of themselves that is more positive and upbeat (Suler 2004). This is often referred to as “emotional labour,” and can be draining and stressful for users who feel pressure to maintain a positive image.

Further, users who express negative emotions online are often met with social disapproval and may be more likely to be ignored, unfriended, or blocked (Kirschner and Karpinski 2010). This is especially true for women, who may be disproportionately targeted by online harassment when they express negative emotions (Hardaker 2015).

Waterloo et al. (2018) mention that " expression of negative emotions was rated as more appropriate for Facebook and Twitter compared to Instagram," meaning Instagram has a social filter according to which positive emotions are more socially acceptable compared to negative ones Waterloo et al. (2018) also show the expression of negative emotions is deemed most suitable on WhatsApp, then Facebook, with Twitter and Instagram following in that order.

The Social Filtering Model can explain this as follows. On WhatsApp we can share personal struggles with one person or with a small group. On Facebook, personal struggles can also be shared in one-to-one communication or in private groups (e.g. depression, mental illness, job loss, death in family, divorce). Google searches would also be an example in which family or friends are not involved, meaning we can search about anything we want using a very low social filter on our negative emotions. However, on Twitter, Instagram, and TikTok, the conversation becomes more open and an increased number of people are watching the content, meaning people tend to share much less information about personal struggles and other negative emotions on these platforms. The Social Filtering Model can thus be used to extend the analysis of Waterloo et al. (2018).

In sum, the online norms for negative emotions on social media involve being respectful and considerate when expressing negative emotions, being open and honest when sharing personal experiences, and being mindful of the platform and community norms. Users who express negative emotions online may be met with social disapproval and are more likely to be ignored, unfriended, or blocked.

Sharing personal struggles online can also be met with social support, but the online norms for negative emotions are typically filtering the content people are publishing about themselves.

7.1.a.2 Social comparisons on social media

Social media platforms such as TikTok, Facebook, Instagram, and Twitter can give individuals a skewed and unrealistic view of others' lives. This phenomenon is also known as "social comparison," where individuals compare their own lives and circumstances to those of others on social media, often leading to feelings of inadequacy and low self-esteem.

Social comparisons on social media refer to the tendency of individuals to compare themselves and their lives to others based on the information and images they see on social media platforms. Research has shown that social comparisons on social media can lead to negative effects on individuals' self-esteem and well-being (e.g. Moreno et al. 2011; Twenge and Campbell 2009). For example, Moreno et al. (2011) show that Facebook use is associated with increased body dissatisfaction and negative body image among adolescent girls, while Twenge and Campbell (2009) prove that increased social comparisons on social media are associated with higher levels of depression and anxiety among college students.

Anderson et al. (2018), Cookingham et al. (2015), Fardouly et al. (2018), and Gündüz (2017) also show that social media provides an unrealistic view of others' lives and often affects negatively peoples' identities and mood.

This unrealistic view of others' lives on social media platforms such as TikTok, Facebook, Instagram, and Twitter is a phenomenon also known as "social media illusion" or "social media façade" (D'Angelo 2019). It refers to the tendency of

people to present a curated and idealised version of themselves online, leading others to believe that their lives are perfect and without problems (Mänty 2019).

Research has proven that this phenomenon is particularly prevalent on platforms such as Instagram, where users are more likely to present a curated and idealised version of themselves (Rosen et al. 2013). Other studies have also found that viewing these curated and idealised versions of others' lives can lead to feelings of envy, low self-esteem, and depression (Tiggemann and Slater 2014).

One study conducted by Garrett (2018), a communication professor at Ohio State University, shows that social media users tend to overestimate the happiness and success of others and underestimate their own happiness and success. This can lead to feelings of inadequacy and dissatisfaction with one's own life.

In this context, upward social comparison on social media refers to the tendency of individuals to compare themselves to others who they perceive as better off in some way, such as having a more attractive appearance, a more successful career, or a more fulfilling lifestyle. This type of comparison can have negative effects on an individual's self-esteem and well-being. Social media platforms such as Facebook, Instagram, and Twitter make it easy for people to make upward comparisons, as they often present highly curated and idealised versions of other people's lives.

Extant research has shown that upward social comparison on social media can lead to feelings of envy and dissatisfaction with one's own life (e.g. Sommers 2012) and that excessive use of social media is associated with increased symptoms of depression and anxiety (e.g. Baumeister 2010).

It is important to note that social comparisons are not always negative, as they can also serve as sources of motivation and inspiration; however, when this practice

becomes excessive or one-sided, it can have negative effects on individual well-being.

Masciantonio et al. (2021) and Verduyn et al. (2017) show that the passive use of Facebook and Twitter is negatively related to well-being due to upward social comparisons. Actively using social network sites can also be positively associated with well-being through social support (Verduyn et al. 2017).

7.1.b The filter bubble algorithms social media sites uses to personalise content based on users' interest

A filter bubble is a phenomenon that occurs when the algorithms used by social media platforms such as Facebook and Google prioritise showing users content that is similar to what they have previously engaged with, rather than showing them a diverse range of content. This can lead to users being presented with a narrow and potentially biased view of the world, as they are not exposed to information that challenges their existing beliefs or attitudes (see e.g. Feezel et al. 2018; Hermida et al. 2012; Spohr et al. 2017).

Filter bubbles are created by the use of personalised algorithms that consider a user's past behaviour and preferences to predict what content they will be most likely to engage with. These algorithms are based on machine learning and use data mining techniques to analyse large amounts of data about users' past behaviours, such as the types of content they have liked, shared, or clicked on.

The concept of filter bubbles was popularised by Eli Pariser in his 2011 book, "The Filter Bubble: What the Internet is Hiding from You." Pariser (2011) argues that these algorithms create a "personal ecosystem of information" that insulates users from diverse perspectives and ideas.

Several studies on the impact of filter bubbles on social media platforms have been

conducted. A 2015 study published in the journal *Science*, Bakshy et al (2015), found that social media users are more likely to be exposed to news stories consistent with their pre-existing beliefs, rather than stories that challenge those beliefs. Another study published in the *Proceedings of the National Academy of Sciences* in 2018, Bail et al (2018) found that users were more likely to click on links that were consistent with their pre-existing beliefs and that this behaviour led to the reinforcement of political polarisation.

It is worth noting that filter bubbles can also be created by the way people choose to interact with social media platforms by seeking out communities and sources that align with their beliefs and avoiding or unfollowing those that do not.

Overall, filter bubbles have been widely criticised for their potential to reinforce existing biases and limit exposure to diverse perspectives.

7.1.c The API filters social media sites use for access to their data

API restrictions on social media refer to the limitations and rules that social media platforms set on the use of their APIs. These restrictions can include limitations on the number of requests that can be made to an API in a given time period, the types of data that can be accessed, and the purposes for which the data can be used.

7.1.c.1 Twitter API

One example of API restriction on social media is Twitter's developer policy, which limits the number of requests that a developer can make to the Twitter API to a certain number during a 15-minute window. The Twitter API has a limit on 0.5 million users per month in Essential Access, 2 million users per month in Elevated Access, 10 million users per month in Academic Research Access, and 10+ million users per month in Enterprise Access. Twitter claims this policy

prevents overloading the Twitter servers and ensures that all developers have fair access to the API. I consider that it is likely part of their business model, where money can buy you access to more data and they maximise the API income source to limit their losses and increase market cap.

Refer to Twitter (2023)

7.1.c.2 Facebook API

Facebook has also several restrictions in place for its API, which they claim are intended to protect user privacy and data security. Some examples of these restrictions are:

Rate limits: Facebook has set limits on the number of requests that can be made to its API within a certain time period. This prevents overloading its servers and protects against malicious or excessive use of the API.

App review: Developers must submit their apps for review before they can access certain features of the Facebook API. This ensures that the app is in compliance with Facebook's policies and terms of service.

User permissions: Facebook's API requires developers to obtain user consent before accessing certain types of data, such as private messages or friend lists. This protects user privacy and ensures that users are aware of what data is being accessed and how it will be used.

Data retention: Facebook's API requires developers to delete user data they no longer need. This protects user privacy and prevents the misuse of data.

Login review: Developers are required to submit their apps for review if they are using Facebook Login, which allows users to sign into an app using their Facebook

credentials. This ensures that the app is in compliance with Facebook's policies and terms of service.

Refer to Facebook (2023).

7.1.c.3 Instagram API

Instagram has implemented API restrictions to limit the amount and type of data that third-party apps can access. Instagram claims that these restrictions were put in place to protect user privacy and prevent misuse of data.

One major restriction is the limitation of the number of requests that third-party apps can make to the Instagram API. According to the Instagram API documentation, "each developer is limited to 5,000 calls per hour per access token" (Instagram 2023). This limitation prevents apps from overloading the Instagram servers and causing performance issues.

Another restriction is the limitation of the data that third-party apps can access. For example, apps are not allowed to access a user's private data, such as their direct messages or personal information (Instagram 2023). Additionally, apps are not allowed to scrape or collect data from Instagram's website, as stated in the Instagram Platform Policy (Instagram 2020).

Finally, Instagram has implemented restrictions on how third-party apps can use the data they access. For example, apps are not allowed to sell or share user data with third parties, as stated in the Instagram Platform Policy (Instagram 2023).

References: Instagram. (2023).

7.1.c.4 TikTok API

The API restrictions on TikTok refer to the limitations set by the company on the use of its application programming interface (API) by third-party developers. TikTok claims that these restrictions protect user data and maintain the integrity of the platform.

According to TikTok's developer documentation, the company's API is only available to authorised partners who have been approved by TikTok to access user data. The API is also subject to strict terms of service and usage guidelines that prohibit developers from using the data for any unauthorised or illegal activities.

Additionally, TikTok has implemented a number of technical restrictions on its API, such as rate limits and IP whitelisting. These measures are designed to prevent unauthorised access and protect against abusing the platform's resources.

In 2019, TikTok introduced new data privacy guidelines for third-party developers, stating that they must "comply with all applicable laws and regulations, including those related to data protection and privacy".

Furthermore, TikTok has restricted the number of requests that developers can make to the platform's API to a maximum of 500 requests per five minutes, which they claim will help to prevent overloading the system with unnecessary requests.

Reference: TikTok (2023)

7.1.c.5 Google Trends API

The API restrictions on Google Trends refer to the limitations and rules set by Google for accessing and using the Google Trends data through their API. Google claims these restrictions ensure that the data are used responsibly and in accordance with Google's terms of service.

According to Google's documentation on the Google Trends API, some of the restrictions include the following. Access to the API is limited to a certain number of requests per day, which varies depending on the type of license to ensure that the data are not overused or abused. The data provided by the API can only be used for non-commercial purposes. This means that they cannot be used for any commercial or business purposes, such as creating products or services for sale. The data provided by the API cannot be used to identify individual users or to track their behaviour to protect the privacy of users and to ensure that the data are used ethically. The data provided by the API cannot be used to create or promote illegal or harmful content. This includes content that is discriminatory, offensive, or promotes hate speech.

Reference: Google (2023)

7.1.c.6 Overall API landscape

In 2018, Instagram and Facebook decreased their API restriction from 5,000 to 200 requests per hour per user. It was the Cambridge Analytica scandal that led to these severe restrictions in API data access for both Instagram and Facebook. This limits the access to analysing big brands on Instagram and Facebook, meaning Google Trends and Twitter now have an advantage in analysing big brand data in terms of social data.

Refer to Facebook (2023) and Techcrunch (2018).

As mentioned above, Google and Twitter data have an advantage in analysing big brands data on social media, compared to Instagram and Facebook. In general, the data API access of Instagram and Facebook restricts the possibilities for predictive analytics with these data and Twitter also has limited data API access compared to Google.
The data API access to Indexed searches on Google, YouTube, Google Images, and Google Shopping, and in some cases actual numbers of searches through Google Ads (formerly Google AdWords), makes the Google data the most unfiltered data in terms of the API data access level.

7.2 Conclusions of the Social Filtering Model

The filtered images of other peoples' lives on social media have been researched and documented for more than a decade, but there is hitherto no model showing the differences in filtering among social media platforms, across both online norms of negative emotions, social comparisons, and API. The filters people use on their own social media content and searches are mainly due to the norms for online negative and positive emotions, and social comparison. This affects individual filter bubbles, which is why filter bubbles are not included in the Social Filtering Model, as it would present some circular causal effects.

The filtering overview provided by the model considers the norms for negative and positive emotions. The filter also covers the unrealistic presentation of peoples' lives on TikTok, Instagram, and to some degree Facebook and Twitter, which is negatively related to well-being due to upward social comparisons. The Social Filtering Model is thus including both the norms for negative and positive emotions, social comparison, and API in a new concept that identifies the filtering

differences among social network sites and web searches.

The Social Filtering Model can be combined with the social data model by Vatrapu et al. (2016). This combined model shows that the Social Filtering Model filters which conversations and interactions we can analyse, depending on which social data platform we have chosen.

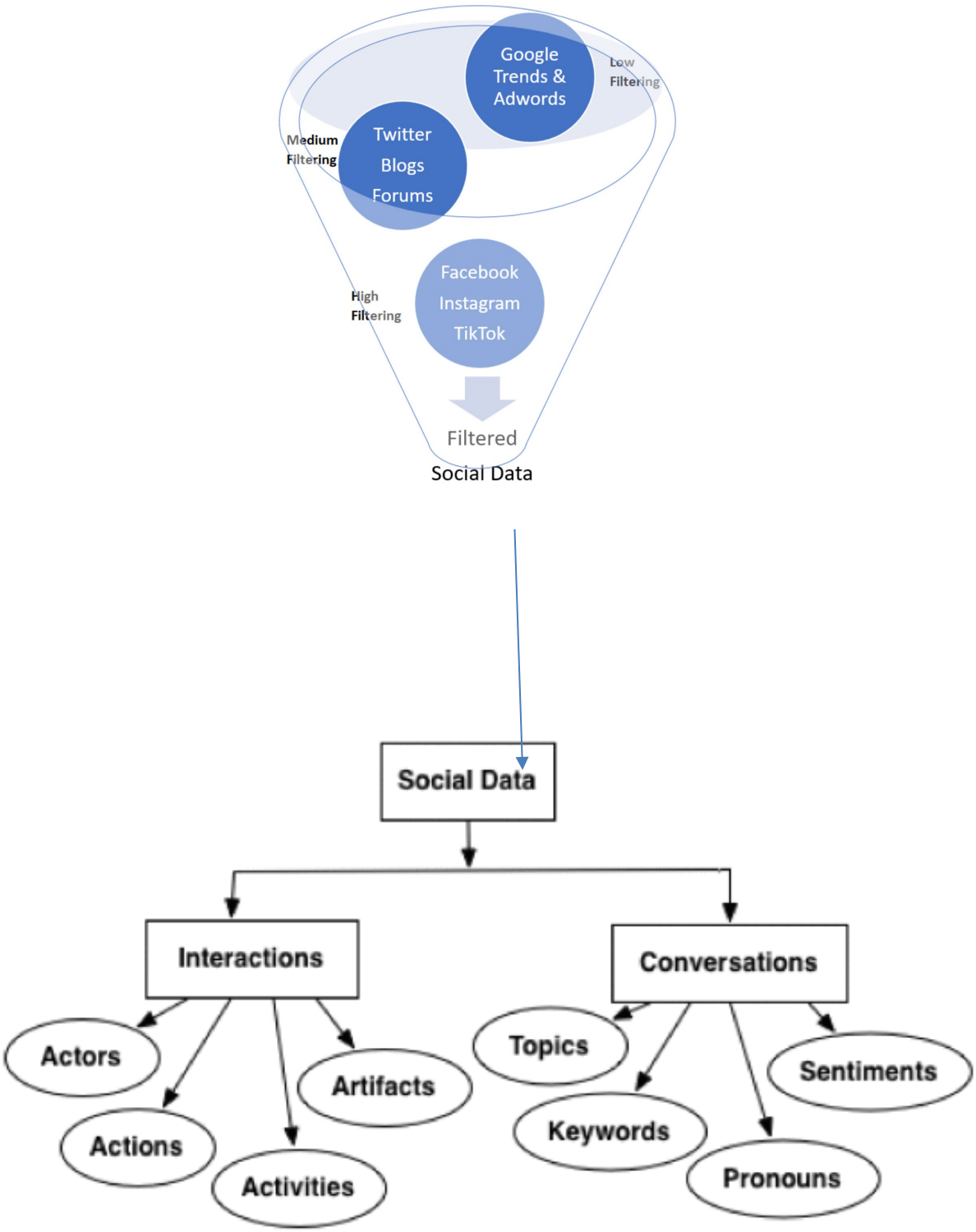


Figure 11: Social Filtering Model by Lassen (2023) combined with the social data model by Vatrapu et al. (2016).

The Social Filtering Model was developed based on the insights gained throughout my PhD research and makes an important contribution to the discussion and

practice of the potential uses of social media, blogs, forums, and web search data. This model is filtering what enters the social data model, depending on which social platform is chosen. The principles of model are used in Predictive Modelling Framework in the following chapter.

8. Predictive Modelling Framework

Model Goal	Stock prices and volatility	Attention Interest Desire	Sales	Trends	Trust Reputation NPS	Opinions	Customer preferences
Relevant Data	Google Trends and Adwords Twitter Blogs Forums StockTwits Influencers	Google Trends and Adwords Twitter Blogs Forums Instagram TikTok Facebook Spotify Influencers YouTube Reviews Interviews	Google Trends and Adwords Twitter Instagram TikTok Facebook Spotify Influencers YouTube	Google Trends and Adwords Twitter Blogs Forums Instagram TikTok Facebook Spotify Influencers YouTube Reviews Interviews	Google Trends and Adwords Twitter Blogs Forums Instagram TikTok Facebook Spotify Influencers YouTube Reviews Interviews	Google Trends and Adwords Twitter Blogs Forums Instagram TikTok Facebook Spotify Influencers YouTube Reviews Interviews	Google Trends and Adwords Twitter Blogs Forums Instagram TikTok Facebook Spotify Influencers YouTube Reviews Interviews
Relevant Methods	Text/topic Mining Statistics Machine learning	Text/topic Mining Statistics Machine learning	Text/topic Mining Statistics Machine learning	Text/topic Mining Visual Analytics Trend-spotting	Sentiment Mining Statistics Machine learning	Opinion Mining Text analytics Trend-spotting	Preference Mining Text analytics Trend-spotting

Table 14, Predictive Modelling Framework

The above model is a more practical version of the predictive model building steps,

presented in Figure 2: Steps for building an empirical model (predictive or explanatory)., based on Shmueli (2011). This new version of the model is designed for social media and web search data and focuses on the goal, data collection, and potential methods in Figure 2.

Table 14 has been expanded into a predictive modelling framework, being inspired by Shmueli (2011) and CRISP-DM, and building on own research and supervision of hundreds of predictive models.

8.1 The steps of the predictive modelling framework

1. Create your research question and identify the key variables. Start by examining the variables that have the most significant impact on the outcome you are trying to predict. Low-to-medium filtered data sources should be included as key variables. These variables will become the foundation of your framework.

2. Define the relationships between variables. Based on the predictive modelling framework in Table 14 and the literature, identify the relationships between the key variables. Determine how they interact with each other and what their relative importance is. Be aware of filtering differences. Low-to-medium filtered data sources will often have higher predictive power and relative importance compared to highly filtered data.

3. Determine the input and output variables. Decide which variables are used as inputs to the framework and which are used as outputs. Input variables will be used to predict the outcome, while the output variables are the predicted outcomes.

4. Organise the variables into a structure. Once you have identified the key

variables, relationships, and inputs and outputs, organise them into a structured framework. Consider using a flowchart, mind map, or other diagram to represent the relationships between variables.

5. Validate the framework. Test the framework using real-world data and evaluate its accuracy out-of-sample or in the future. Make any necessary adjustments to improve prediction accuracy.

6. Incorporate data sources and algorithms. Decide which data sources and algorithms will be used to feed into the framework and make predictions. Ensure that these data sources and algorithms are consistent with the framework's structure.

7. Refine the framework. Continuously evaluate and refine the framework as more data become available and as new insights emerge, as data relations change over time. Make updates to improve the accuracy of predictions and ensure that the framework remains relevant.

8.2 Guidelines for size of dataset, in the Predictive Modelling Framework

The brands relevant for modelling using social data in Table 14 need to have a certain size in terms of social media and web searches. For example, the Twitter data for iPhones in Paper I and the Facebook data for H&M in Paper II, were large enough to model sales. However, the Facebook and Google search data for Mikkeller beer in Paper III were not large enough to model sales, although Mikkeller is a very trendy and popular brand. Therefore, brands need to have a relatively large size on social media and web searches for having a modelling potential using the above framework. It is always possible to model opinions and customer preferences, even for smaller brands, but all other dimensions in the

framework only work for brands with large data amounts in terms of social media and web searches. The threshold for data size is often determined experimentally, but the rough guidelines are over 50k Tweets, Facebook reactions, or Google searches mentioning the brand. The brands modelled in Papers I and II had over 500 million tweets and 15 million Facebook reactions, respectively, while the Apple stock symbol in Paper V had millions of Google searches.

The main differences in the above data types is how transparent people are on the different platforms. Google searches are closer to the truth, as we can search for anything we want, while platforms such as Facebook, Instagram, and TikTok present a highly polished picture of peoples' lives, as these platforms favours success, which is measured by likes and other user actions. The Social Filtering Model in Table 11 shows and explains these differences and determines the potential use of social data in the Predictive Modelling Framework.

The stock market has been modelled in many articles using Twitter and Google Trends data as inputs, showing good predictive power for these data types. Refer to Bollen et al. 2011; Zhang et al. 2011; Pagolu et al. 2016; Hu et al. 2018; Bijl et al. 2016; Batra et al. 2018; Oliveira et al. 2013; Oh et al. 2011; Preis et al. 2013), Nguyen et al (2015), Curme et al (2014).

Purchase intentions, which include attention, interest, and desire, have been modelled in many articles using Twitter, Google Trends, Instagram, Facebook, TikTok, and social media influencer data as input, with good predictive power for these data types (see e.g. Astuti et al. 2018; Zhang et al. 2021; Bag et al. 2019; Lo et al. 2016; Guo et al. 2019; Erkan et al. 2016; Lim et al. 2017; Hermanda et al. 2019).

Sales have been modelled in many articles using Twitter, Google Trends, Instagram, Facebook, blogs, and social media influencer data as input, also with

good predictive power for these data types. See e.g. Brown et al. 2013; Lassen et al. 2014, 2016; Asur et al. 2010, Zhang et al (2017), Fan et al (2017), Hasan et al (2018), Kim et al (2019), Liang et al (2015) and Lee et al (2019).

The predictive modelling of sales also works best with low-to-medium filtered data sources, although some models have worked relatively well with highly filtered data sources like Facebook and Instagram.

Google Trends data work better for the predictive modelling of sales because most people use Google searches in their customer journey before buying a product and they search with almost no filtering.

However, TikTok, Facebook, and Instagram data are rather used to present products in filtered glamorous settings, and these data do not often reflect the unfiltered honest opinions of users about these products.

Trends have been modelled in many articles using Twitter, Google Trends, Instagram, Facebook, blogs and influencers social media data as input data, finding good predictive power for these data types (see e.g. Granata et al. 2019; Gloor 2017; Altshuler et al. 2012).

Trust, reputation, and Net Promoter Scores have been modelled in many articles using Twitter, Google Trends, Instagram, Facebook, blogs, and social media influencer data as input data, with good predictive power for these data types (see e.g. Khadangi et al. 2013; Kandias et al. 2013; Prada et al. 2020; Peetz et al. 2016; DuBois et al. 2011; Vedula et al. 2017; Zaki et al. 2016; van Velthoven 2014; Pop et al. 2021).

Opinions and customer preferences have been modelled in many articles using Twitter, Google Trends, Instagram, Facebook, blogs, and social media influencer

data as inputs, with good predictive power for these data types (see e.g. Jiang et al. 2019; Sobkowicz et al. 2012; Fang et al. 2020; Nguyen et al. 2012). The opinions and preferences of customers can be used for product and service innovation and development by simply considering customers' comments, likes, dislikes, and wishes for features.

8.2 Digital maturity impacts the use of social data

Digital maturity refers to the ability of an organization to use digital technologies to change traditional ways of doing business, improve performance, and deliver value to customers. It encompasses the integration of digital technologies into the organization's strategy, operations, culture, and customer engagement. Digital maturity have an impact on the use of social media, blogs, forums and web search data.

See e.g. Ransbotham et al (2015), Hull et al (2020), Hanna et al (2011) and Nambisan et al (2017).

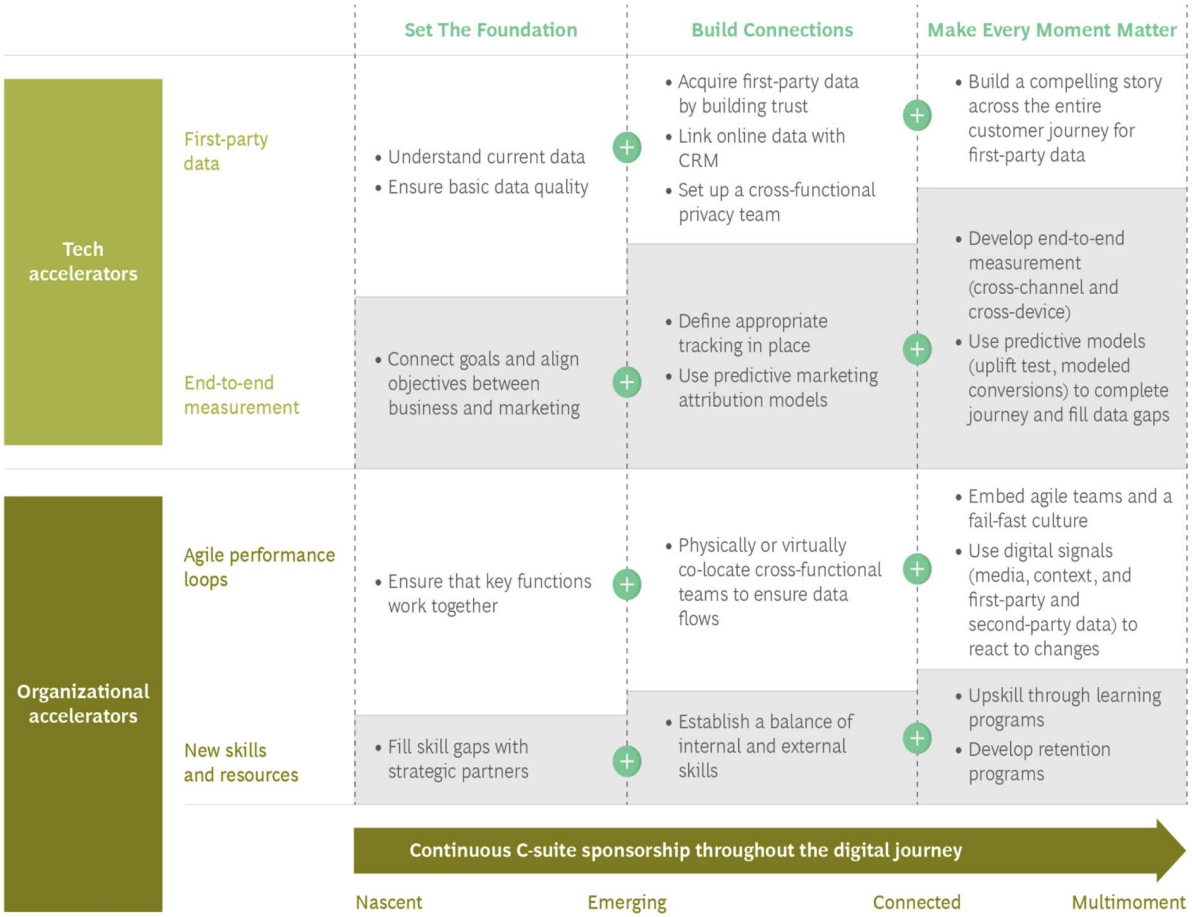


Figure 12, Digital Maturity Model, by Boston Consulting Group (2021)

Organisations should be aware of their level of digital maturity, as this can limit their potential use of social media, blogs, forums and web search data in the Table 14, Predictive Modelling Framework. Digital maturity analysis can address problems such as aversion, (Kane et al (2015)), opacity (Westerman et al (2011)), and human-machine collaboration (Brynjolfsson & McAfee (2014)).

For an academic version of the digital maturity model, I refer to Lasrado et al. (2016), who focus on the use of social media and size of IT investments and measure digital maturity based on this use. Instead, the digital maturity model of the Boston Consulting Group (2021) focuses more on how social media is used and measures digital maturity stages based on this. Their model also includes intelligent links between social media, customer journey, and the use of first- and

second-party data. I consider that the use of social media data needs to be strategically aligned with CRM, customer journey, and first-, second-, and third-party data throughout all relevant departments in an organisation. In this sense, the model of the Boston Consulting Group (2021) is the most inclusive and pragmatic one. However, it is not enough to focus on how much organisations use social media, but it is also important to focus on how social media is used and aligns with the CRM, customer journey, and strategy of the company.

Teichert (2019) examines 22 digital maturity models over 2011–2018, 12 by academics and 10 by practitioners. His conclusion is that most models provide an incomplete picture of digital maturity and that the description of digital maturity stages is inconsistent across models. Teichert (2019) also mentions that the manufacturing domain is overrepresented in the examined maturity models, while the service domain, for example, requires more research and focus.

The Boston Consulting Group's (2021) digital maturity model is one of the newest, considering that the digital landscape moves fast. Their model also has some of the problems pointed out by Teichert (2019), but considers the shift from third-party to first- and second-party data, which was a major disruptive factor in advertising and marketing in 2021. Specifically, Apple, Microsoft, and Mozilla have stopped access to third party cookie data in their browsers in 2021 and Google will also follow this trend in 2023. This has fundamentally changed the data landscape of digital maturity models (see BBC 2021).

9. Findings

Chapter 1 presented the research questions that this PhD thesis is aiming to answer,

also describing the links of the research questions with the five papers and the key contributions. Below, I detail the findings based on the research questions.

RQ1: Which social data types can be used to predict consumer purchase behaviours and to what extent does it work for different brand types?

Paper I showed it is possible to predict iPhone sales with Twitter data and paper II showed we can predict H&M sales with Facebook data. Conversely, paper III demonstrated it is not possible to predict Mikkeller beer sales with either Google or Facebook data. Research into the reasons for these predictive modelling successes and failure has shown important insights about both the nature and size of the social data, which are determining the potential use of social data in predictive models. These insights led to the development of the Figure 10: Social Filtering Model. This model shows how filtering is determining which content we can consider and measure on different social platforms. This model is the most important contribution of this PhD thesis, as it lays the foundation for a more scientific discussion on the potential uses for social data. The Social Filtering Model also led to the development of Table 14, Predictive Modelling Framework. This model is considering both the filtering and size of data, to make recommendations about which social data to use for specific modelling contexts. The degree of filtering on social data platforms is determining the potential use of these data for sales modelling. The size of social data is determining which brands are suitable for sales modelling with social data.

RQ2: What, if any, are the explanatory mechanisms for social data based predictive models for consumer purchase behaviours?

Customer journey models are used in the first four papers of this PhD thesis, as the conceptual model explaining how all social data can be placed in one of the phases of a customer journey model. As social data are a proxy for the activities in each phase of a customer journey model, the social data acting as proxies for the last two phases of a customer journey model contains strong links between the behaviour regarding social data and purchasing behaviour. Therefore, the small proportion of all social data belonging to the last two phases in the customer journey model is the reason why some social data can predict sales if the dataset is big enough. Refer to the size recommendations in chapter 8.2 Guidelines for size of dataset, in the Predictive Modelling Framework.

This is best illustrated using a simple customer journey model, as in below Figure 12.

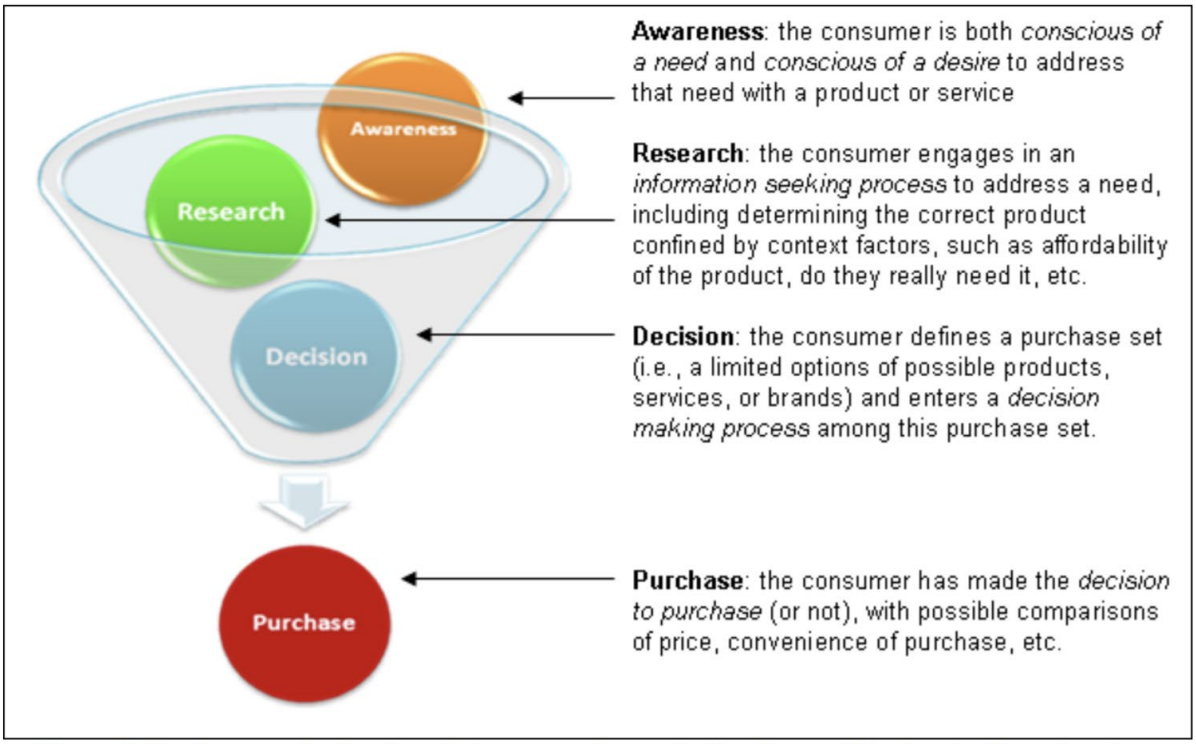


Figure 13: Awareness, research, decision, and purchase journey model.

Source: Jansen et al. (2011). Copyright permission from Dr. Jim Jansen

The 500+ million tweets from the iPhone model in Paper I, could be used as an example where about 20% of the tweets belong to the decision phase and about 10% to the purchase phase. These numbers are only examples of what the actual proportions could be. This example illustrates that if these proportions are relatively stable in the iPhone Twitter dataset, then a simple regression model can learn the links between the tweets containing 'iPhone' belonging to the decision and purchase phases and the actual iPhone sales.

Customer journey models have become increasingly complex over time, but the older and simpler customer journey models illustrate the clear relationships between social data and sales.

RQ3: To what extent can social data provide predictors for investor behaviour?

Paper V showed the predictive power of Google searches for Apple stock volatility applied to Apple stock investor behaviour, based on the customer journey concepts developed in papers I–IV. The Apple stock and Google search data were chosen for this predictive modelling case based on both the Social Filtering Model and the Predictive Modelling Framework. Choosing the Apple stock with one of the largest web search volumes ensured a large dataset, while the choice of Google searches ensured a very low social filter on these data. The model failed in modelling the Apple stock price, but visual graphs of Google searches for the Amazon stock symbol showed a higher correlation with the Amazon stock price than for the Apple stock. Therefore, the method in paper V can be successfully used modelling stock prices for stocks other than Apple. Paper V also showed the stock symbols for big tech stocks being the most important Google searches for stock volatility, and this led to an important distinction between private and professional investors' use of stock symbols on Google searches. In other words, Google searches have good predictive power for at least stock volatility, if the data are large enough. Chapter 8. Predictive Modelling Framework showed the successful use of both Google searches and Twitter data for the predictive modelling of investor behaviour in many articles. As such, it would be interesting to combine Twitter and Google search data for the predictive modelling of investor behaviour. The main contribution of paper V is the investor journey model, explaining why the Google search data have predictive power for investor behaviour. Another contribution is the identification of private and professional investors' different uses of Google searches, and the high importance of stock symbols, among all the stock related Google searches for predictive modelling of investor behaviour.

RQ4: How can extant social data models be adapted to better inform the predictive models of consumer and investor behaviours?

Figure 10: Social Filtering Model., shows important differences in the nature of

social data and how it affects which conversations we listen to and measure on different social platforms. Therefore, this model is an important contribution for the scientific discussion of how social data can be used. An example is social media platform TikTok, which started in 2016, but commenced its high growth journey when the COVID-19 pandemic started in the early 2020.

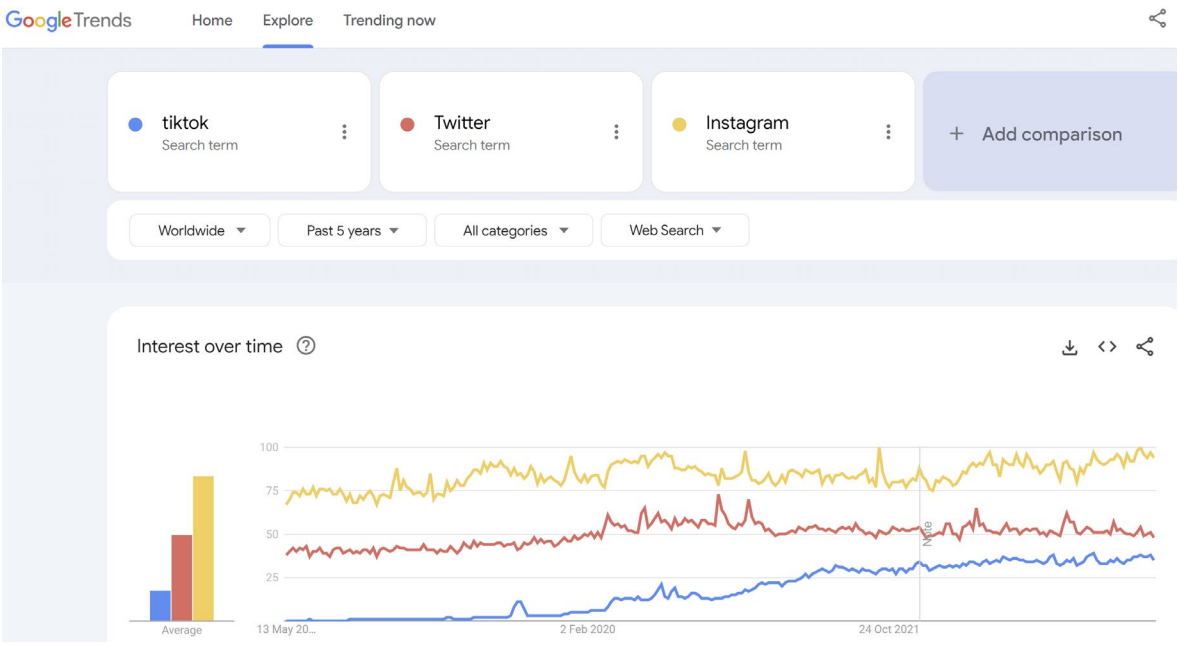


Figure 14: Google searches for TikTok, Twitter & Instagram 2018–2023.

Source:https://trends.google.com/trends/explore?date=today%205-y&q=tiktok,Twitter,Instagram

According to above Figure 14, in April 2023, TikTok had 75% of the Google search volume of Twitter and almost 40% of the Google search volume of Instagram.

A customer journey model, using data from Google searches, blogs, forums, Instagram, Facebook, and TikTok has to consider the differences in the nature of all these data sources. The Social Filtering Model explains some important differences in these data sources, which will bring better insights for using them. TikTok proved the perfect entertainment medium over the COVID-19 isolation

period, but from a filtering sense, it is fair to ask what can we use these data for? Many trends start on TikTok, making TikTok difficult to overlook for predicting trends, but trends also start on YouTube, Instagram, and other platforms. TikTok features videos from 15–180 seconds, so it is like the original Twitter concept with 140-character tweets blended with YouTube. Only time will show if TikTok also increases the length of their videos, similar to Twitter´s expansion to 280 characters tweets in November 2017. As of now, TikTok has a high filtering score in the Social Filtering Model, together with Instagram and Facebook, because the concept is very liked and follower driven, similar to Instagram and Facebook – the entertainment factor on TikTok is measured in likes, shares, and followers. However, while Facebook and Instagram favours success in peoples' lives, TikTok is favouring a successful entertainment factor. Twitter has more dimensions as a microblog and social network, covering other topics than TikTok, Instagram, and Facebook. The format on Twitter is rawer, as spelling errors and typos are more socially acceptable, whereas immediacy and being first with news are often more important. The likes, replies, retweets, and shares on Twitter are still ensuring a medium filtering in the Social Filtering Model, but the data have more potential uses compared to the highly filtered data from TikTok, Instagram, and Facebook. Searches on Google and YouTube have low filtering in the Social Filtering Model, as people can search for anything they want. Social filters are only applied if someone is watching the search or if people are worried their employer is monitoring their Internet access at work. This makes search data from Google and YouTube raw and unfiltered and explains why data from Google Trends are often outperforming social data from more filtered platforms in many models, and also can be considered as more truthful in many cases.

This is also shown in Figure 10: Social Filtering Model., which was used to develop chapter 8. Predictive Modelling Framework.

10. Conclusions

This thesis explained why social media and web search data have predictive power for sales and stock price volatility if social data are big enough. Specifically, I demonstrated the successful use of Twitter and Facebook as predictors for iPhone and H&M sales and gave the example of Mikkeller for social data being too small to have predictive power for sales. I also demonstrated the successful use of Google search data as predictors for Apple stock price volatility. In this final chapter, I discuss the contributions of the results of this PhD thesis to the literature, their implications for practice, and my future research strategy.

10.1 Contributions to the Literature

The initial contribution of this doctoral dissertation is explaining why social media and web search data have predictive power for sales and stock price volatility, while the main theoretical contributions are (i) the development of predictive sales models using Twitter and Facebook data and (ii) the development of a predictive model for stock price volatility using Google search data.

In particular, this thesis conceptualizes the customer journey as the main model explaining why social media data have predictive power for sales. Further, by building on the customer journey model, it develops the investor journey model as the main model for explaining why web search data have predictive power for stock price volatility.

From a practical viewpoint, the thesis also offers detailed guidelines for researchers and practitioners to implement this conceptual model to the predictive modelling of sales, stock price volatility, and other domains. In paper IV, a more generalized

predictive model for several domains is presented and the logic of the customer journey model is used as an example in epidemiology. Specifically, in epidemiology, all social media texts and flu-related web searches can be categorized into the different phases of spread, incubation, immunity, resistance, susceptibility etc. This is the conceptual model-based explanation for why social media and web search data can predict the spreading of the seasonal flu, COVID-19, and many other infectious diseases if the data are large enough.

A second contribution is testing of how large social media and web search data have to be to predict sales. See chapter 8.2 Guidelines for size of dataset, in the Predictive Modelling Framework. Papers I and II showed that Twitter and Facebook data for iPhone and H&M were large enough to predict sales, while paper III demonstrated that the social media and web search data for Mikkeller were not. A lab test of Google search data for Nike also showed these data were large enough to predict Nike worldwide sales out-of-sample on a quarterly basis. Based on these experiences, Apple stocks were chosen for modelling in paper V, as Google search data needed to be the largest possible.

A third contribution of this thesis is identifying four different cases of social-data-based prediction models for smartphones, clothing, beer, and stocks. One refers to Twitter data predicting iPhone sales (paper I), a second one to Facebook data predicting H&M sales (paper II), and a third one to social media and web search data failing to predict Mikkeller Brewery sales due to the small size of the Mikkeller brand and the related social data. The fourth case is demonstrating Google search data can be used as predictors for Apple stock price volatility, succeeding in reasonably forecasting out-of-sample, with a mean average prediction error of 38.5% (paper V). In paper IV, the cases from papers I and II are reviewed together with 38 research articles using social media and web search data for predictive modelling in five domains: epidemiology, sociopolitics, culture,

marketing, and finance/economics.

A fourth contribution is the introduction of chapter 7. Social Filtering Model, which explains important social filtering differences for social media, blogs, forums and web search data. This can start a scientific discussion about the potential uses of social data.

A fifth contribution is the introduction of chapter 8. Predictive Modelling Framework, which builds on the Social Filtering Model, and gives practical guidelines for the use of social media, blogs, forums and web search data, in the domain of marketing and finance predictive modelling.

10.2 Managerial Implications

The traditional data used for forecasting within the sales and marketing include mostly CRM data, qualitative interview data, and market research reports. One example of classical sales forecasting methods can be found at https://corporatefinanceinstitute.com/resources/knowledge/modeling/forecasting-methods/ (retrieved 26 October 2021).

The use of social media and web search data for forecasting purposes within sales and marketing is increasing but it is not wide enough just yet. However, as described by Cui et al. (2018), social media data can significantly improve the accuracy of existing sales forecasting methods, based on their experiment involving an online clothing company and its Facebook data.

Pepsi and Procter & Gamble are examples of large brands that have adopted social media data in forecasting demand early on. Pepsi has worked with marketing prediction agency Black Swan since 2013 for predicting both demand trends and also later sales (Stewart 2018).

Pepsi has also worked from 2020 with Tastewise agency, who claim to have monitored 95 million menu items, 226 billion recipe interactions, and 22.5 billion social posts, among other consumer touchpoints. Pepsi also has their own demand accelerator AI solution built on internal datasets, which includes data from more than 100 million U.S. households, with first-party data at the individual level representing more than half (Lazzaro 2021).

By employing more social media and web search data for forecasting sales and trends for services and products, large and medium brands can better adapt to a fast-changing world. Further, by employing the techniques for predictive analytics in this thesis, brands can change their digital maturity.

10.3 Limitations

Twitter, Facebook, Google search and YouTube Search data were tested in this PhD,
as predictors for consumer purchase and investor behaviour. The insights from these data were used to develop Figure 10: Social Filtering Model and Table 14, Predictive Modelling Framework.

These two new social data models could have been further substantiated by using more social data. The reason for not including more social data in the five papers, was the limited access to other data sources. In addition to the data in the five papers, I also used social data from Roskilde Festival and industrial projects to further develop the two new social data models.

Field research from Roskilde Festival Big Data Lab 2015-2018 on Twitter, Facebook, Instagram, Spotify, blogs, forums, Google search and YouTube search data, and participation in industrial influencer projects with Instagram and TikTok data, were also included in the development of Figure 10: Social Filtering Model and Table 14, Predictive Modelling Framework

Conceptual limitations were only to focus on customer journey models, Social Network Analysis and Social Set Analysis. The scope of this PhD thesis could have been expanded to include conceptual social data models with, for example, Big 5 personality types and Big 6 human emotions. As shown by Alan et al (2016) and Azucar et al (2018), personality types are determining how humans act on social media. So using conceptual models with these dimensions could have revealed more patterns on human purchase and investor behaviour and could also have further substantiated Figure 10: Social Filtering Model. The Social Filtering Model is also broadly generalizing into low, medium and high filtering for only Google search, Twitter, blogs, forums, Facebook, Instagram and TikTok data. The Social Filtering Model will be developed into a more operational model, which can measure the filtering more precise per actor group and on more social data types.

A methodological limitation is the lack of interview data. The touch points in both customer and investor journeys could have been explored in new directions with interview data. As mentioned below in subsection 10.4 Future Research Directions , interview data can reach dimensions not seen by the big social data from social media and web searches.

A social influencer for the iPhone sales is Justin Bieber. His influence on iPhone sales could be modelled together with other iPhone influencers, as a part of the

customer journey using SNA. However, for modelling total iPhone sales, hundreds of millions social media posts containing 'iPhone' are relevant, which was shown in paper I, but tracing all these social media posts to influencers is not useful for total sales modelling purposes. Using Justin Bieber as an example of an iPhone sales influencer, there is no doubt he has contributed making the iPhone cooler and trendier among his fans, and also changed the iPhone's reputation among them. As such, for modelling marketing variables as brand coolness, trendiness, and reputation, social influencers are important. Such modelling can be done with both SNA and SSA, where each social influencer is a subset in the total dataset. For the iPhone sales prediction in paper I of this PhD, more than 500 million tweets containing 'iPhone' were used, but social influencers were not analysed because of the limited scope of the paper. However, social influencers could have been analysed and modelled with the same Twitter dataset, for modelling sales and other marketing variables such as brand coolness, trendiness, and reputation.

In other words, influencers can be analysed and modelled using SSA or set analysis rather than SNA (see e.g. del Fresno et al. 2016; Rios et al. 2019). Chapter 4.5.1 Social Network Analysis vs Social Set Analysis, also lists relevant use of influencers under the frameworks of both SSA and set analysis.

Methodological limitations were also to only use regression and time series methods for the first three papers in this PhD. Paper IV analyse statistical and machine learning predictive models using social data. Lasso and Autometrics regression were the only machine learning models used in paper V. 40 new statistical and machine learning predictive models are applied on datasets from papers I, II, III and V in chapter 6. iPhone, H&M, Mikkeller and Apple datasets, in the light of 40 new models.
But more machine learning models fine tuned and tested more in depth, could

have found new important predictors and data patterns. As pointed out by Breiman (2001), the classical statistical modelling approaches should be supplemented with algorithmic approaches, which is missing in the first three papers of this PhD. Paper IV analyses algorithmic and statistical approaches on 38 predictive models, and paper V is testing Lasso and Autometrics regression.

As mentioned in chapter 8.2 Digital maturity impacts the use of social data, organisations should be aware of their level of digital maturity, as this can limit their potential use of social media, blogs, forums and web search data. Digital maturity analysis can address problems such as aversion, (Kane et al (2015)), opacity (Westerman et al (2011)), and human-machine collaboration (Brynjolfsson & McAfee (2014)).

10.4 Future Research Directions

I was able to observe where quantitative big data analytics methods could supplement QRM methods and vice versa in my research group at the Center for Business Data Analytics at CBS (see http://bda.cbs.dk/). Therefore, my future research on social-data-based predictive models will include mixed methods, as purely quantitative methods can miss important human behaviours, as can purely qualitative methods. An example is a two-hour interview that can reach deep levels of values and visions, which quantitative methods can supplement and explore further. Figure 10: Social Filtering Model will be developed into a more operational model, which can measure the filtering more precise per actor group and on more social data types.

References

Abboud, L., 2019. Big brands turn to big data to rekindle growth. Financial Times, 26 August 2019. https://www.ft.com/content/abc231ac-c288-11e9-a8e9-296ca66511c9 (retrieved 20 December 2021).

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research, 25(3), 443-448.

Alan, A.K. and Kabadayı, E.T., 2016. The effect of personal factors on social media usage of young consumers. Procedia-Social and Behavioral Sciences, 235, pp.595-602.

Altshuler, Y., Pan, W. and Pentland, A.S., 2012, April. Trends prediction using social diffusion models. In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 97–104). Springer, Berlin, Heidelberg.

Anderson, M. and Jiang, J., 2018. Teens, social media & technology 2018. Pew Research Center, 31(2018), pp. 1673–1689.

AppleInsider, 2015. Apple shuts down Topsy two years after acquisition. Forwards Webpage to iOS 9 Search Support. https://appleinsider.com/articles/15/12/15/apple-shuts-down-topsy-two-years-after-acquisition-forwards-webpage-to-ios-9-search-support (retrieved 12 October 2020).

Aslam, B., Jun, C. S., & Karim, A. (2020). Sentiment analysis of social media big data using machine learning techniques. IEEE Access, 28890-28907.

Astuti, B. and Putri, A.P., 2018. Analysis on the effect of Instagram use on consumer purchase intensity. Review of Integrative Business and Economics Research, 7, pp. 24–38.

Asur, S. and Huberman, B.A., 2010. Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 31 August–3 September, Toronto, ON, Canada.

Azucar, D., Marengo, D. and Settanni, M., 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and individual differences, 124, pp.150-159.

Bag, S., Tiwari, M.K. and Chan, F.T., 2019. Predicting the consumer's purchase intention of durable goods: An attribute-level analysis. Journal of Business Research, 94, pp. 408–419.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences, 115(37), 9216-9221. DOI: https://doi.org/10.1073/pnas.1804840115

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. Science, 348(6239), 1130-1132. DOI: https://doi.org/10.1126/science.aaa1160

Bapna, R., Gupta, A., Ray, G., & Singh, V. K. (2020). Digital first: The ontological inversion in digital strategy research. MIS Quarterly, 44(1), 261-286.

Baronchelli, Andrea, et al. "The emergence of consensus: Opinion formation on social networks." Proceedings of the National Academy of Sciences 111.29 (2014): 10779-10784.

Batra, R. and Daudpota, S.M., 2018, March. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1–5). IEEE.

Baumeister, R. F. (2010). Is there a downside to social comparison? Current Directions in Psychological Science, p. 81-85.

BBC, June 2021. Google tracking cookies ban delayed until 2023. https://www.bbc.com/news/technology-57611701 (retrieved 3 January 2022).

Bifet, A., Holmes, G., Pfahringer, B., & Gavaldà, R. (2010). "Mining Adaptive Micro-Clusters from Data Streams using Ensemble Methods". In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/1835804.1835934

Bijl, L., Kringhaug, G., Molnár, P. and Sandvik, E., 2016. Google searches and stock returns. International Review of Financial Analysis, 45, pp.150–156.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp.1–8.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. DOI: https://doi.org/10.1016/j.jocs.2010.12.007

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. Science, 323(5916), 892-895.

Boston Consulting Group, 2021. The fast track to digital marketing maturity. https://www.bcg.com/publications/2021/the-fast-track-to-digital-marketing-maturity
(retrieved 20 November 2021).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Springer, Heidelberg.

Breiman, L. (1996). "Bagging Predictors". Machine Learning, 24(2), 123-140. doi:10.1023/A:1018054314350

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), pp.199-231.

Brown, D. and Fiorella, S., 2013. Influence marketing: How to create, manage, and measure brand influencers in social media marketing. Que Publishing.

Brown, Z. and Tiggemann, M., 2016. Attractive celebrity and peer images on Instagram: effect on women's mood and body image. Body Image, volume 19, 2016, pp. 37–43.

Brynjolfsson, E., & McAfee, A. (2014). The Second Machine Age: Work,

Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton & Company.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning (pp. 108-122).

Bzdok, D., Altman, N. and Krzywinski, M., 2018. Statistics versus machine learning. Nature Methods 15, pp. 233–234.

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (2nd ed.). New York: Cambridge University Press.

Chaffey, D., & Ellis-Chadwick, F. (2019). Digital marketing: Strategy, implementation, and practice. Pearson UK.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165-1188.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, Y., & Xie, J. (2018). Online consumer review, product variety, and the long tail effect. Information Systems Research, 29(2), 328-341).

Choi, H. and Varian, H., 2012. Predicting the present with Google Trends. Economic Record 88, pp. 2–9.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), 2-9. DOI: https://doi.org/10.1111/j.1475-4932.2012.00809.x

Chou, H. T. G., & Edge, N. (2012). "They are happier and having better lives than I am": The impact of using Facebook on perceptions of others' lives. Cyberpsychology, Behavior, and Social Networking, 15(2), 117-121. DOI: https://doi.org/10.1089/cyber.2011.0324

Christoffersen, B., Matin, R. and Mølgaard, P., 2019. Can machine learning models capture correlations in corporate distresses?. SSRN Electronic Journal. https://www.nationalbanken.dk/da/publikationer/Documents/2018/10/WP_128_Oktober.pdf (retrieved 31 October 2021).

Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (2007). Data mining methods for knowledge discovery. Springer Science & Business Media.

Clements, M. and Hendry, D., 2002. An overview of economic forecasting. In A companion to economic forecasting. Oxford: Blackwell, pp. 118.

Cookingham, L.M. and Ryan, G.L., 2015. The impact of social media on the sexual and social wellness of adolescents. Journal of Pediatric and Adolescent Gynecology, 28(1), pp. 2–5.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

Creswell, J.W., 2009. Mapping the field of mixed methods research. Qualitative Methods, 3(2), pp. 95–108.

Crotty, M., 1998. Foundations of social research: Meaning and perspective in the research process. Sage.

Cui, R., Gallino, S., Moreno, A. and Zhang, D.J., 2018. The operational value of social media information. Production and Operations Management, 27(10), pp. 1749–1769.

Curme, C., Preis, T., Stanley, H. E., & Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. Proceedings of the National Academy of Sciences, 111(32), 11600-11605. DOI: https://doi.org/10.1073/pnas.1324054111

D'Urso, J., 2020. How the coronavirus pandemic is changing social media. https://reutersinstitute.politics.ox.ac.uk/news/how-coronavirus-pandemic-changing-social-media (retrieved 31 October 2021).

D'Angelo, P. (2019). The Social Media Illusion: Understanding the Impact of Social Media on Mental Health. Journal of Technology in Human Services, 37(4), 2019, pp 1-9.

Danish Loyalty Clubs conference 2016, held at 29 January 2016 at CBS. Conference presentation page 70-72 contains Net Promoter Score model. http://efficiens.nu/dl/de-danske-loyalitetsprogrammer.pdf (retrieved 6 December 2021)

Davenport, T.H., & Harris, J.G. (2007). Competing on Analytics: The New Science of Winning. Boston: Harvard Business School Press.

del Fresno Garcia, M., Daly, A.J. and Segado Sanchez-Cabezudo, S., 2016. Identifying the new Influences in the Internet Era: Social Media and Social Network Analysis. Revista Española de Investigaciones Sociológicas, 153.

Dewey, J., 1923. Democracy and education: An introduction to the philosophy of education. Macmillan.

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry (pp. 88-121). Oxford University Press.

DuBois, T., Golbeck, J. and Srinivasan, A., 2011, October. Predicting trust and distrust in social networks. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing (pp. 418–424). IEEE.

Durbin, J., & Koopman, S. J. (2012). Time Series Analysis by State Space Methods (2nd ed.). Oxford University Press.

Edelman, D. C., & Singer, M. (2015). Competing on customer journeys. Harvard Business Review, 93(11), 88-100.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32(2), 407-499.

Erkan, I. and Evans, C., 2016. The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. Computers in Human Behavior, 61, pp.47–55.

Facebook (2023)
https://developers.facebook.com/docs/facebook-login/review
https://developers.facebook.com/docs/graph-api/overview/rate-limiting/
https://developers.facebook.com/docs/instagram-basic-display-api/overview
https://developers.facebook.com/docs/marketing-apis/policies#app_review

https://developers.facebook.com/docs/marketing-apis/policies#data_retention
https://developers.facebook.com/docs/marketing-apis/policies#rate_limiting
https://developers.facebook.com/docs/marketing-apis/policies#user_permissions

Fan, L., & Li, H. (2017). The impact of Instagram influencer marketing on firm sales. Journal of Marketing Research, 54(6), 745-758.

Fang, Y., Chen, X., Song, Z., Wang, T. and Cao, Y., 2020. Modelling propagation of public opinions on microblogging big data using sentiment analysis and compartmental models. In Natural Language Processing: Concepts, Methodologies, Tools, and Applications (pp. 939–956). IGI Global.

Fardouly, J. and Holland, E., 2018. Social media is not real life: The effect of attaching disclaimer-type labels to idealized social media images on women's body image and mood. New Media & Society, 20(11), pp. 4311–4328.

Fardouly, J., & Vartanian, L. R. (2016). Social media and body image concerns: Current research and future directions. Current Opinion in Psychology, 9, 1-5. DOI: https://doi.org/10.1016/j.copsyc.2015.09.005

Fardouly, J., Diedrichs, P. C., Vartanian, L. R., & Halliwell, E. (2015). Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. Body Image, 13, 38-45. DOI: https://doi.org/10.1016/j.bodyim.2014.12.002

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Advances in knowledge discovery and data mining (pp. 1-34). AAAI Press.

Feehan, B., 2020. The impact of coronavirus on social media engagement for brands. https://www.rivaliq.com/blog/coronavirus-on-social-media-engagement-for-brands/
(retrieved 7 September 2020).

Feezell, J.T., 2018. Agenda setting through social media: The importance of incidental news exposure and social filtering in the digital era. Political Research Quarterly, 71(2) pp.482-494.

Financial Times, 2021. Critics raise alarm over Big Tech's most powerful tools. https://www.ft.com/content/d9d505fe-d1f6-4a0d-9eae-c6b208f72cee

Fischler, M.A., Bolles, R.C. (1981). "Random Sample Consensus: A Paradigm for

Model Fitting with Applications to Image Analysis and Automated Cartography". Communications of the ACM, 24(6), 381–395. DOI: 10.1145/358669.358692

Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas, Project Number 21-49-004, Report Number 4.

Freund, Y., & Schapire, R. E. (1997). "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". Journal of Computer and System Sciences, 55(1), 119-139. doi:10.1006/jcss.1997.1504

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 29(5), 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, p 137-144.

Garcia, A. & Mirra, N. (2019). The different ways we express negative emotions on Instagram and Twitter. Journal of Computer-Mediated Communication. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jcc4.12317

Garrett, R. K. (2018). Social media use and perceptions of social well-being. PloS one, e0191841.

Gartner, 2020. Predictive modeling. https://www.gartner.com/en/information-technology/glossary/predictive-modeling (retrieved 21 August 2020).

Géron, A. (2019). "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" (2nd ed.). O'Reilly Media.

Gloor, P.A., 2017. Sociometrics and human relationships: Analyzing social networks to manage brands, predict trends, and improve organizational performance. Emerald Group Publishing.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science, 333(6051), 1878-1881. DOI: https://doi.org/10.1126/science.1202775

Google (2023)

https://developers.google.com/trends/v1/docs/faq
https://developers.google.com/trends/v1/guides/usage_limits
https://developers.google.com/trends/v1/terms

Granata, G., Moretta Tartaglione, A. and Tsiakis, T., eds., 2019. Predicting trends and building strategies for consumer engagement in retail environments. IGI Global.

Gündüz, U., 2017. The effect of social media on identity construction. Mediterranean Journal of Social Sciences, 8(5), p. 85.

Guo, L., Hua, L., Jia, R., Zhao, B., Wang, X. and Cui, B., 2019, July. Buying or browsing?: Predicting real-time purchasing intent using attention-based deep network with multiple behavior. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1984–1992).

Gupta, M., & Bhatia, V. (2014). Predictive analytics in healthcare: A review. Journal of Healthcare Engineering, pp. 457-478.

Haferkamp, N., & Krämer, N. C. (2011). Social comparison 2.0: Examining the effects of online profiles on social-networking sites. Cyberpsychology, Behavior, and Social Networking, 14(5), 309-314. DOI: https://doi.org/10.1089/cyber.2010.0120

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. Business Horizons, 54(3), 265-273.

Hardaker, C. (2015). Online abuse of women in the political sphere. Journal of Language Aggression and Conflict, 3(1), 1-22.

Harvey, A. C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Hasan, M. M., & Kim, H. (2018). The impact of Facebook marketing on firm sales. Journal of Marketing Research, 55(2), 165-180.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (2nd ed.). Springer.

Hendry, D. F., & Doornik, J. A. (2014). Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics. MIT Press.

Hendry, D. F., & Krolzig, H.-M. (2005). The Properties of Automatic Gets Modelling. The Economic Journal, 115(502), C32-C61.

Hermanda, A., Sumarwan, U. and Tinaprillia, N., 2019. The effect of social media influencer on brand image, self-concept, and purchase intention. Journal of Consumer Sciences, 4(2), pp.76–89.

Hermida, A., Fletcher, F., Korell, D. and Logan, D., 2012. Share, like, recommend: Decoding the social media news consumer. Journalism studies, 13(5-6), pp.815-824.

Hochman, N., & Schwartz, R. (2012). Visualizing Instagram: Tracing cultural visual rhythms. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012), Dublin, Ireland, 881-884.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Holland, G. and Tiggemann, M. (2016) A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes. Body Image, 17, pp. 100–110.
Hu, H., Tang, L., Zhang, S. and Wang, H., 2018. Predicting the direction of stock markets using optimized neural networks with Google Trends. Neurocomputing, 285, pp.188–195.

Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73-101.

Hübner, A., Holzapfel, A., & Kuhn, H. (2020). The impact of digital technologies on customer journey touchpoints and satisfaction. Journal of Retailing and Consumer Services, 57, 102258.

Hull, G., & Dougherty, I. (2020). Digital Marketing Strategy: An Integrated Approach to Online Marketing. Kogan Page Publishers.

IBM. 2020. What is predictive analytics?. https://www.ibm.com/analytics/predictive-analytics (retrieved 21 August 2020).

Instagram. (2023) https://www.instagram.com/about/legal/terms/platform/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer.

Jank, W. and Shmueli, G., 2010. Modeling online auctions. New York: John Wiley & Sons.

Jansen, J. and Schuster, S., 2011. Bidding on the buying funnel for sponsored search and keyword advertising. Journal of Electronic Commerce Research, 12.

Jiang, H., Kwong, C.K., Kremer, G.O. and Park, W.Y., 2019. Dynamic modelling of customer preferences for product design using DENFIS and opinion mining. Advanced Engineering Informatics, 42, 100969.

Jordan, M.I., & Mitchell, T.M. (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, p. 255-260.

Jorgensen, B. (1987). Exponential dispersion models. Journal of the Royal Statistical Society: Series B (Methodological), 49(2), 127-162.

Kalampokis, E., Tambouris, E. and Tarabanis, K., 2013. Understanding the predictive power of social media. Internet Research 23(5), pp. 544–559.

Kallas, P., 2019. Top 15 Most popular social networking sites and apps [2021],. https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/ (retrieved 16 June 2020).

Kandias, M., Stavrou, V., Bozovic, N., Mitrou, L. and Gritzalis, D., 2013, December. Can we trust this user? Predicting insider's attitude via YouTube usage profiling. In 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing (pp. 347-354). IEEE.

Kane, G. C., Palmer, D., Phillips, A. N., Kiron, D., & Buckley, N. (2015).

Strategy, not technology, drives digital transformation. MIT Sloan Management Review and Deloitte University Press.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3149-3157).

Kelleher, C. D., Mac Namee, B., & Morley, M. (2015). Using Twitter sentiment analysis to predict stock market patterns. Decision Support Systems, 69, 21-30.

Khadangi, E. and Bagheri, A., 2013, October. Comparing MLP, SVM and KNN for predicting trust between users in Facebook. In ICCKE 2013 (pp. 466-470). IEEE.

Kim, H., & Hasan, M. M. (2019). The impact of blog marketing on firm sales. Journal of Marketing Research, 56(1), 81-96.

Kirschner, P. A., & Karpinski, A. C. (2010). Facebook and academic performance. Computers in Human Behavior, 26(6), 1237-1245.

Koehrsen, W., 2019. Thoughts on the two cultures of statistical modeling. https://towardsdatascience.com/thoughts-on-the-two-cultures-of-statistical-modeling-72d75a9e06c2 (retrieved 31 October 2021).

Kuhn, T.S. (1962). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

la Cour, Lisbeth, Milhøj, A., Vatrapu, R., and Lassen, N. B. 2018. Predicting the daily sales of Mikkeller bars using Facebook data. Symposium i Anvendt Statistik, 22–24 January 2018.

Lasrado, L., Vatrapu, R. and Andersen, K.N., 2016. A set theoretical approach to maturity models: Guidelines and demonstration. ICIS 2016 Proceedings. Association for Information Systems. AIS Electronic Library (AISeL), Atlanta, GA, Proceedings of the International Conference on Information Systems, vol. 37, p. 20, 2016 International Conference on Information Systems, ICIS 2016, Dublin,Ireland,11/12/2016. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1355&context=icis2016

Lassen, N.B., 2022. “Google searches linked to apple stock volatility ups and downs. Proceedings of the 43rd Symposium i Anvendt Statistik, August 2022, Denmark.

Lassen, N.B., la Cour, L. and Vatrapu, R., 2017. Predictive analytics with social media data, in Sloan L., and Quan-Haase A., eds. The SAGE handbook of social media research methods, Sage, pp. 328–341.

Lassen, N.B., la Cour, L., Milhøj, A. and Vatrapu, R. 2017. "Social media data as predictors of Mikkeller sales?. Proceedings of the 39th Symposium i Anvendt Statistik, University of Southern Denmark, Odense, Denmark, 23–24 January 2017.

Lassen, N.B., Madsen, R. and Vatrapu, R. 2014. Predicting iPhone sales from iPhone tweets. Proceedings of the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, 1–5 September 2014, Ulm, Germany.

Lassen, N.B., Vatrapu, R., la Cour, L., Madsen, R. and Hussain, A. 2016. Towards a theory of social data: Predictive analytics in the era of big social data. Proceedings of the 38th Symposium i Anvendt Statistik, Copenhagen Business School, Frederiksberg, Denmark, 25–27 January 2016.

Lavidge, R.J. and Steiner, G.A., 1961. "A model for predictive measures of advertising effectiveness. Journal of Marketing, pp. 59–62.

Lazzaro, S., 2021. How PepsiCo uses AI to create products consumers don't know they want. https://venturebeat.com/2021/06/28/how-pepsico-uses-ai-to-create-products-consumers-dont-know-they-want/ (retrieved 27 October 2021).

Lee, Y., & Kim, H. (2019). The impact of social media influencer marketing on firm sales. Journal of Marketing Research, 57(1), 121-136.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. Journal of Marketing, 80(6), 69-96.

Liang, X., Sun, L., & Zhang, Y. (2015). Predicting sales with social media data. Journal of Marketing Research, 52(4), 497-509.

Lim, X.J., Radzol, A.M., Cheah, J. and Wong, M.W., 2017. The impact of social media influencers on purchase intention and the mediation effect of customer attitude. Asian Journal of Business Research, 7(2), pp.19–36.

Lo, C., Frankowski, D. and Leskovec, J., 2016, August. Understanding behaviors that lead to purchasing: A case study of Pinterest. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining

(pp. 531–540).

Makridakis, S., Spiliotis, E. and Assimakopoulos, V., 2020. The M4 competition: 100,000 time series and 61 forecasting methods," International Journal of Forecasting, 36, pp. 54–¬74.

Mänty, J. (2019). The Social Media Illusion: How Social Media Platforms Affect Our Understanding of Reality. Journal of Media Psychology, 31(3), 113-125.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. In Data Mining and Knowledge Discovery in Real Life Applications (pp. 438-453). I-Tech Education and Publishing. DOI: 10.5772/6438.

Masciantonio, A., Bourguignon, D., Bouchat, P., Balty, M. and Rimé, B., 2021. Don't put all social network sites in one basket: Facebook, Instagram, Twitter, TikTok, and their relations with well-being during the COVID-19 pandemic. PloS One, p.e0248384.

McCullagh, P., & Nelder, J.A. (1989). Generalized Linear Models (2nd ed.). London: Chapman and Hall/CRC.

McKinsey, 2009. The customer decision journey, https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey (retrieved 18 April 2021).

Merton, R.K. (1973). The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 3111-3119.

Moreno, M. A., Jelenchick, L. A., Cox, E., Young, H. N., & Walker, D. L. (2011). Problematic Internet use and psychosocial well-being among US college students. Cyberpsychology, Behavior, and Social Networking, 14(5), 183–189. https://doi.org/10.1089/cyber.2009.0411

Morselli, D., & Bruggeman, J. (2018). Social Set Analysis. In D. Morselli (Ed.), Social Networks, Terrorism and Counter-terrorism: Radical and Connected (pp. 15-32). Routledge.

Mukkamala, R.R., Hussain, A. and Vatrapu, R. 2014. Towards a formal model of

social data. 3rd International Congress on Big Data (IEEE Big Data 2014), June 2014.

Nambisan, S., Lyytinen, K., Majchrzak, A., & Song, M. (2017). Digital Innovation Management: Reinventing innovation management research in a digital world. MIS Quarterly, 41(1).

Nguyen, L.T., Wu, P., Chan, W., Peng, W. and Zhang, Y., 2012, August. Predicting collective sentiment dynamics from time-series social media. In Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (pp. 1-8).

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611. DOI: https://doi.org/10.1016/j.eswa.2015.07.052

Oh, C. and Sheng, O., 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. Presented at ICIS 2011 conference. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.651.1224&rep=rep1&type=pdf

Oliveira, N., Cortez, P. and Areal, N., 2013, September. On the predictability of stock market behavior using StockTwits sentiment and posting volume. In Portuguese conference on artificial intelligence (pp. 355-365). Springer, Berlin, Heidelberg.

Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B., 2016, October. Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

Papacharissi, Z. (2010). A private sphere: Democracy in a digital age. Polity Press.

Pariser, Eli. The filter bubble: what the Internet is hiding from you. Penguin Press, 2011.

Patel, S., 2015. The research paradigm – Methodology, epistemology and ontology – Explained in simple language. http://salmapatel.co.uk/academia/the-research-paradigm-methodology-epistemology-and-ontology-explained-in-simple-language/
(retrieved 16 September 2020).

Pati, Y. C., Rezaiifar, R., & Krishnaprasad, P. S. (1993). Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, 1, 40-44.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Peetz, M.H., de Rijke, M. and Kaptein, R., 2016. Estimating reputation polarity on microblog posts. Information Processing & Management, 52(2), pp.193-216.

Penguin, 2011. -Bakshy, Eytan, et al. "Exposure to ideologically diverse news and opinion on Facebook." Science 348.6239 (2015): 1130-1132.

Perrin, A. (2016). Social media usage: 2005-2015. Pew Research Center. Retrieved from https://www.pewresearch.org/internet/2016/11/11/social-media-usage-2005-2015/

Pop, R.A., Săplăcan, Z., Dabija, D.C. and Alt, M.A., 2021. The impact of social media influencers on travel decisions: The role of trust in consumer decision journey. Current Issues in Tourism, pp. 1–21.
Popper, K. (1959). The Logic of Scientific Discovery. London: Hutchinson.

Porter, T.M. (1986). The Rise of Statistical Thinking, 1820-1900. Princeton: Princeton University Press.

Prada, A. and Iglesias, C.A., 2020. Predicting reputation in the sharing economy with Twitter social data. Applied Sciences, 10(8), p. 2881.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific Reports, 3, 1684. DOI: https://doi.org/10.1038/srep01684

Ransbotham, S., Kiron, D., & Prentice, P. K. (2015). The talent dividend. MIT

Sloan Management Review, 56(4), 1-17.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse Additive Models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(5), 1009-1030.

Rios, S.A., Aguilera, F., Nuñez-Gonzalez, J.D. and Graña, M., 2019. Semantically enhanced network analysis for influencer identification in online social networks. Neurocomputing, 326, pp.71–81.

Ritter, B. et al. (2016). Social media and the well-being of individuals with mental health conditions. Journal of Medical Internet Research. Retrieved from https://www.jmir.org/2016/11/e342/

Rosen, L. D., Whaling, K., Rab, S., Carrier, L. M., & Cheever, N. A. (2013). Is Facebook creating "iDisorder"? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. Computers in Human Behavior, 29(3), 1243-1254.

Roth, C., Herzog, M., & Bandyopadhyay, T. (2016). Social set analysis: A set-theoretical approach to the study of social groups. Social Networks, 44, 157-168.

Rousidis, D., Koukaras, P. and Tjortjis, C., 2019. Social media prediction: A literature review. Multimedia Tools and Applications, 79 (9–10), pp. 1–33.

Sabater, J. and Sierra, C., 2002, July. Reputation and social network analysis in multi-agent systems. In Proceedings of the first international joint conference on Autonomous agents and multiagent systems: Part 1 (pp. 475-482).

Samet, A., 2020. How the coronavirus is changing US social media usage. https://www.emarketer.com/content/how-coronavirus-changing-us-social-media-usage
(retrieved 4 September 2020).
SAS Institute, 2020. What is predictive analytics?. https://www.sas.com/ko_kr/insights/analytics/predictive-modeling-techniques.html
(retrieved 29 July 2020).

Shah, S. (2020). LazyPredict: A Python library for lazy machine learning model comparison. GitHub. Retrieved from https://github.com/shankarpandala/lazypredict

Sharma, P. (2020). LazyPredict: Python library to Compare Various Machine Learning Models. Retrieved from https://www.geeksforgeeks.org/lazypredict-python-library-to-compare-various-machine-learning-models/

Sheedy, C., 2019. https://www.intheblack.com/articles/2019/12/01/what-is-social-network-analysis

Shmueli, G. and Koppius, O., 2011. Predictive analytics in information systems research. MIS Quarterly, 35, pp. 553–572.

Shmueli, G., 2010. To Explain or to Predict?. Statistical Science, 25(3), pp. 289–310.

Singh, R. (2021). An Introduction to Lazy Predict: A Python Library for Lazy People. Retrieved from https://www.analyticsinsight.net/an-introduction-to-lazy-predict-a-python-library-for-lazy-people/

Sobkowicz, P., Kaschesky, M. and Bouchard, G., 2012. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. Government Information Quarterly, 29(4), pp. 470-479.

Sommers, M. S. (2012). Social comparison and social media: The impact of Facebook on young women's body image concerns. Sex Roles, p. 363-377.

Spohr, D., 2017. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. Business information review, 34(3), 150-160.

St. Elmo Lewis, E., 1899. Side talks about advertising. The Western Druggist, 21, pp. 65–66.

Statista, 2018. Global Apple iPhone sales from 3rd quarter 2007 to 4th quarter 2018 (in million units). https://www.statista.com/statistics/263401/global-apple-iphone-sales-since-3rd-quarter-2007/ (retrieved 20 September 2020).

Statista, 2023. Most popular social networks worldwide as of January 2023, Ranked by number of active users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (retrieved 12 May 2023).

Stern, S.E., Gregor, S., Martin, M.A., Goode, S. and Rolfe, J., 2004. A Classification tree analysis of broadband adoption in Australian households. Proceedings of the 6th International Conference on Electronic Commerce, Delft,

The Netherlands, 25–27 October, pp. 451–456.

Stewart, R., 2018. PepsiCo uses data science to decide its next crisp flavour, now it could inform its marketing. https://www.thedrum.com/news/2018/04/17/pepsico-uses-data-science-decide-its-next-crisp-flavour-now-it-could-inform-its (retrieved 27 October 2021).

Suler, J. (2004). The online disinhibition effect. Cyberpsychology & Behavior, 7(3), 321-326.

Techcrunch (2018) https://techcrunch.com/2018/07/02/facebook-rolls-out-more-api-restrictions-and-shutdowns/

Teichert, R., 2019. Digital transformation maturity: A systematic review of literature. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 67(6), pp. 1673–1687.

Teradata, 2020. What are predictive analytics?. https://www.teradata.co.uk/Glossary/What-are-Predictive-Analytics (retrieved 26 August 2020).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Tiggemann, M., & Slater, A. (2014). NetGirls: The Internet, Facebook, and the Secret Lives of Teenagers. New York: Henry Holt and Company.

TikTok (2023)
https://developers.tiktok.com/docs/guidelines-for-developers
https://developers.tiktok.com/docs/requests-and-responses

Trainor, K. J., Andzulis, J. M., Rapp, A., & Agnihotri, R. (2014). Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. Journal of Business Research, 67(6), 1201-1208.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM 2014), Ann Arbor, MI, USA, 505-514.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. Statistics: Applications and New Directions. Proceedings of

the Indian Statistical Institute Golden Jubilee International Conference (pp. 579-604).

Twenge, J. M., & Campbell, W. K. (2009). The narcissism epidemic: Living in the age of entitlement. Free Press.

Twitter (2023)
https://developer.twitter.com/en/developer-terms/policy
https://developer.twitter.com/en/docs/twitter-api/tweet-caps

van Velthoven, S.T.M., 2014. Sentiment analysis on social media to predict Net Promoter Score. (Master's thesis, Eindhoven University of Technology)
https://pure.tue.nl/ws/files/46988096/783003-1.pdf

Varian, H.R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, vol. 28, no. 2, p. 3-28.

Varun, S. (2020). Lazy Predict: Fit and evaluate all the models from scikit-learn with a single line of code. Retrieved from https://towardsdatascience.com/lazy-predict-fit-and-evaluate-all-the-models-from-scikit-learn-with-a-single-line-of-code-5e63b5f2c0e1

Vatrapu, R., Mukkamala, R.R., Hussain, A. and Flesch, B., 2016. Social set analysis: A set theoretical approach to big data analytics. IEEE Access, 4, pp. 2542–2571.

Vedula, N., Parthasarathy, S. and Shalin, V.L., 2017, June. Predicting trust relations within a social network: A case study on emergency response. In Proceedings of the 2017 ACM on Web Science Conference (pp. 53-62).

Venkatesh, V., Brown, S., and Bala, H. 2013. "Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems," MIS Quarterly: Management Information Systems (37), pp. 21–54.

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J. and Kross, E., 2017. Do social network sites enhance or undermine subjective well-being? A critical review. Social Issues and Policy Review, 11.1 (2017): 274-302.

Vogel, E. A., Rose, J. P., Roberts, L. R., & Eckles, K. (2014). Social comparison, social media, and self-esteem. Psychology of Popular Media Culture, 3(4), 206-222. DOI: https://doi.org/10.1037/ppm0000047

Voortman, M.C., 2015. Validity and reliability of web search based predictions for car sales (Master's thesis, University of Twente).

Vosoughi, S., Deb Roy, A., & Roy, D. (2017). The spread of true and false news online. Science, 359(6380), 1146-1151.

Wang, D., Shi, X., McFarland, D. A., & Leskovec, J. (2016). Measurement error in network data: A re-classification. Social Networks, 47, 99-112.

Wang, T-W.. Rees, J. and Kannan, K.N., 2008. The association between the disclosure and the realization of information security risk. Working Paper, Purdue University.

Waskom, M. (2021). LazyPredict: A Comprehensive Tool for Comparing Machine Learning Models. Retrieved from https://www.analyticsvidhya.com/blog/2021/01/lazypredict-a-comprehensive-tool-for-comparing-machine-learning-models/

Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press.

Waterloo, S.F., Baumgartner, S.E., Peter, J. and Valkenburg, P.M., 2018. Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. New Media & Society, 20(5), pp. 1679–2096.

Wells, G., Horowitz, J. and Seetharaman, D., 2021. Facebook knows Instagram is toxic for teen girls, company documents show. Wall Street Journal, 14 September 2021.

Westerman, G., Calméjane, C., Bonnet, D., Ferraris, P., & McAfee, A. (2011). Digital Transformation: A Roadmap for Billion-Dollar Organizations. MIT Center for Digital Business and Capgemini Consulting.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39).

Wu, X., & Zhang, C. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, pp. 97-107.

Zaki, M., Kandeil, D., Neely, A. and McColl-Kennedy, J.R., 2016. The fallacy of

the net promoter score: Customer loyalty predictive model. Cambridge Service Alliance, 10, pp.1-25.

Zaltman, G., 1996. Metaphorically speaking. Marketing Research, 8(2), p.13.

Zeng, X., Zhu, H., & Chen, Z. (2017). A set-theoretical approach to feature-based social media data analysis. In International Conference on Web Information Systems Engineering (pp. 61-76). Springer, Cham.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the twenty-first international conference on Machine learning (p. 116). ACM.

Zhang, W., Daim, T. and Zhang, W., 2021. Investigating consumer purchase intention in online social media marketing: A case study of Tiktok. Available at SSRN 3971795.

Zhang, X., Fuehres, H. and Gloor, P.A., 2011. Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia-Social and Behavioral Sciences, 26, pp.55-62.

Zhang, Y., Sun, L., & Srinivasan, P. (2017). Predicting sales with Google Trends. Journal of Business Research, 79, 68-76.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265-286.

Østergaard Jacobsen, P., ed., 2020. CRM 5.0 – De ustyrlige kunder i en digital tidsalder: Mindset, strategi, ledelse og performance i fremtidens forretningsmodeller. Efficiens, Rungsted Kyst.

Appendices

Appendix 1, python LazyPredict code for paper I, predicting iPhone sales with Twitter data

[1]

```
#importing necessary libraries
import pandas as pd
import seaborn as sns
%matplotlib inline
import statistics as st
import matplotlib.pyplot as plt
from numpy import mean
from numpy import std
```

[2]

```
#importing the iphone dataset as a dataframe
iphone=pd.read_csv('iPhone.csv')
iphone.head()
iphone.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62 entries, 0 to 61
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Period     62 non-null     object
 1   Quarter    62 non-null     int64
 2   Year       62 non-null     int64
 3   Sales      62 non-null     float64
 4   Tweets     19 non-null     float64
 5   Sentiment  19 non-null     float64
dtypes: float64(3), int64(2), object(1)
memory usage: 3.0+ KB
```

[3]

```
pip install lazypredict
```

[4]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
from sklearn.metrics import mean_absolute_error
```

[5]

```
#Set Variables (X) and Target (Y) . Remove  'Period' as that is the same as year and Quarter and remove Sales from X.
#Divide data into train/test
#Fit the model
X =iphone.drop(["Sales", 'Period'], axis=1)
Y = iphone["Sales"]
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2, random_state = 64)
reg = LazyRegressor(verbose=0, ignore_warnings=False, custom_metric=mean_absolute_error)
models,predictions = reg.fit(X_train, X_test, y_train, y_test)
models
```

```
100%|██████████| 41/41 [00:01<00:00, 27.39it/s]
```

LazyPredict Results table is documented in
Table 10, LazyPredict results for paper I, predicting iPhone sales with Twitter data

Appendix 2, python LazyPredict code for paper II, predicting H&M sales with Facebook data

[1]

```
#importing necessary libraries
import pandas as pd
import seaborn as sns
%matplotlib inline
import statistics as st
import matplotlib.pyplot as plt
from numpy import mean
from numpy import std
import datetime
```

[2]

```
pip install lazypredict
```

[3]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
```

[4]

```
#importing the hm dataset as a dataframe
hm = pd.read_csv('H&M.csv')
hm.head()
hm.info()
hm
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Year         24 non-null     float64
 1   Quarter      24 non-null     float64
 2   FBLikes      24 non-null     float64
 3   Sales        24 non-null     float64
 4   Unnamed: 4   0 non-null      float64
 5   Unnamed: 5   0 non-null      float64
 6   Unnamed: 6   0 non-null      float64
 7   Unnamed: 7   0 non-null      float64
 8   Unnamed: 8   0 non-null      float64
 9   Unnamed: 9   0 non-null      float64
 10  Unnamed: 10  0 non-null      float64
 11  Unnamed: 11  0 non-null      float64
 12  Unnamed: 12  0 non-null      float64
dtypes: float64(13)
memory usage: 2.7 KB
```

:

	Year	Quarter	FBLikes	Sales
0	2009.00	1.00	6.00	27282.00
1	2009.00	2.00	37053.00	31070.00
2	2009.00	3.00	25845.00	27587.00
3	2009.00	4.00	43523.00	32758.00
4	2010.00	1.00	32434.00	29095.00
5	2010.00	2.00	46942.00	31604.00
6	2010.00	3.00	76010.00	31475.00
7	2010.00	4.00	130316.00	34792.00
8	2011.00	1.00	101261.00	28708.00
9	2011.00	2.00	122256.00	32499.00
10	2011.00	3.00	84763.00	31511.00
11	2011.00	4.00	161618.00	36191.00
12	2012.00	1.00	215388.00	32503.00
13	2012.00	2.00	151985.00	36947.00
14	2012.00	3.00	853013.00	33568.00
15	2012.00	4.00	1656204.00	37930.00
16	2013.00	1.00	1395677.00	33146.00
17	2013.00	2.00	1895557.00	36923.00
18	2013.00	3.00	749560.00	37411.00
19	2013.00	4.00	1872211.00	36495.00
20	2014.00	1.00	1018929.00	37524.00
21	2014.00	2.00	1383915.00	44181.00
22	2014.00	3.00	839567.00	45259.00
23	2014.00	4.00	371768.00	43000.00

[5]

```
#clean the dataset and convert Year and Quarter to int
hm_cleaned =hm.loc[:,["Year", 'Quarter', "FBLikes", "Sales"]]
hm_cleaned = hm_cleaned.drop(24)
hm_cleaned['Year'] = hm_cleaned['Year'].astype('int64')
hm_cleaned['Quarter'] = hm_cleaned['Quarter'].astype('int64')
hm_cleaned
```

	Year	Quarter	FBLikes	Sales
0	2009	1	6.00	27282.00
1	2009	2	37053.00	31070.00
2	2009	3	25845.00	27587.00
3	2009	4	43523.00	32758.00
4	2010	1	32434.00	29095.00
5	2010	2	46942.00	31604.00
6	2010	3	76010.00	31475.00
7	2010	4	130316.00	34792.00
8	2011	1	101261.00	28708.00
9	2011	2	122256.00	32499.00
10	2011	3	84763.00	31511.00
11	2011	4	161618.00	36191.00
12	2012	1	215388.00	32503.00
13	2012	2	151985.00	36947.00
14	2012	3	853013.00	33568.00
15	2012	4	1656204.00	37930.00
16	2013	1	1395677.00	33146.00
17	2013	2	1895557.00	36923.00

18	2013	3	749560.00	37411.00
19	2013	4	1872211.00	36495.00
20	2014	1	1018929.00	37524.00
21	2014	2	1383915.00	44181.00
22	2014	3	839567.00	45259.00
23	2014	4	371768.00	43000.00

[6]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
from sklearn.metrics import mean_absolute_error
```

[7]

```
#Set Variables (X) and Target (Y)
#Divide data into train/test
#Fit the model
X =hm_cleaned.drop(["Sales"], axis=1)
Y = hm_cleaned["Sales"]
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.35, random_state = 64)
reg = LazyRegressor(verbose=0, ignore_warnings=False, custom_metric= mean_absolute_error)
models,pred = reg.fit(X_train, X_test, y_train, y_test)
models
```

```
100%|██████████| 41/41 [00:03<00:00, 13.22it/s]
```

LazyPredict Results table is documented in Table 11, LazyPredict results for paper II, predicting H&M sales with Facebook data

Appendix 3, python LazyPredict code for paper III, predicting Mikkeller beer sales with Google searches

[1]

```
#importing necessary libraries
import pandas as pd
import seaborn as sns
%matplotlib inline
import statistics as st
import matplotlib.pyplot as plt
from numpy import mean
from numpy import std
import datetime
```

[2]

```
pip install lazypredict
```

[3]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
```

[4]

```
#importing the Mikkeller dataset as a dataframe
mikkeller = pd.read_csv('mikkeller.csv')
mikkeller.head()
mikkeller.info()
mikkeller
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 33 entries, 0 to 32
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Month           33 non-null     object
 1   Sales           33 non-null     float64
 2   Google          33 non-null     int64
 3   YouTube         33 non-null     int64
 4   GoogleShopping  33 non-null     int64
 5   Unnamed: 5      0 non-null      float64
 6   Unnamed: 6      0 non-null      float64
 7   Unnamed: 7      0 non-null      float64
 8   Unnamed: 8      0 non-null      float64
 9   Unnamed: 9      0 non-null      float64
 10  Unnamed: 10     0 non-null      float64
dtypes: float64(7), int64(3), object(1)
memory usage: 3.0+ KB
```

[5]

```
#Clean the dataset
mikkeller_cleaned =mikkeller.loc[:,["Month", 'Sales', "Google", "YouTube", "GoogleShopping"]]
mikkeller_cleaned
```

	Month	Sales	Google	YouTube	GoogleShopping
0	Jan-14	20.83	76	23	31
1	Feb-14	25.76	82	23	20
2	Mar-14	30.30	90	20	24
3	Apr-14	22.09	97	23	13
4	May-14	52.81	91	23	31
5	Jun-14	21.10	85	78	28
6	Jul-14	32.70	76	48	31
7	Aug-14	42.36	85	36	34
8	Sep-14	61.04	85	26	26
9	Oct-14	48.87	82	33	25
10	Nov-14	38.32	89	75	25
11	Dec-14	20.47	92	50	31
12	Jan-15	55.92	90	40	12
13	Feb-15	33.21	93	51	25
14	Mar-15	43.23	98	52	40
15	Apr-15	59.63	100	43	35
16	May-15	45.74	88	41	22
17	Jun-15	66.22	84	46	17
18	Jul-15	58.45	87	31	17
19	Aug-15	77.32	89	42	11
20	Sep-15	85.97	81	44	11

21	Oct-15	55.97	94	19	10
22	Nov-15	49.27	89	60	32
23	Dec-15	34.58	100	100	16
24	Jan-16	48.47	87	20	46
25	Feb-16	56.82	93	20	21
26	Mar-16	58.27	90	26	43
27	Apr-16	72.98	98	20	40
28	May-16	69.47	89	16	38
29	Jun-16	51.81	88	26	30
30	Jul-16	43.37	89	36	20
31	Aug-16	99.23	88	18	34
32	Sep-16	71.59	90	18	14

[6]

```
#Extract Year and Months
# Define a function to extract yYear and Months as integers
def extract_year_month(date_string):
    date_obj = datetime.datetime.strptime(date_string, '%b-%y')
    year = date_obj.year
    month = date_obj.month
    return (year, month)

# Apply the function to the 'Month' column and create new columns for yYear and Months
mikkeller_cleaned[['Year', 'Months']] =
mikkeller_cleaned['Month'].apply(extract_year_month).apply(pd.Series)

#Drop the Column Month
mikkeller_cleaned =  mikkeller_cleaned.drop(["Month"], axis=1)
mikkeller_cleaned
```

	Sales	Google	YouTube	GoogleShopping	Year	Months
0	20.83	76	23	31	2014	1
1	25.76	82	23	20	2014	2
2	30.30	90	20	24	2014	3
3	22.09	97	23	13	2014	4
4	52.81	91	23	31	2014	5
5	21.10	85	78	28	2014	6
6	32.70	76	48	31	2014	7
7	42.36	85	36	34	2014	8
8	61.04	85	26	26	2014	9
9	48.87	82	33	25	2014	10
10	38.32	89	75	25	2014	11
11	20.47	92	50	31	2014	12
12	55.92	90	40	12	2015	1
13	33.21	93	51	25	2015	2
14	43.23	98	52	40	2015	3
15	59.63	100	43	35	2015	4
16	45.74	88	41	22	2015	5
17	66.22	84	46	17	2015	6
18	58.45	87	31	17	2015	7
19	77.32	89	42	11	2015	8
20	85.97	81	44	11	2015	9

21	55.97	94	19	10	2015	10
22	49.27	89	60	32	2015	11
23	34.58	100	100	16	2015	12
24	48.47	87	20	46	2016	1
25	56.82	93	20	21	2016	2
26	58.27	90	26	43	2016	3
27	72.98	98	20	40	2016	4
28	69.47	89	16	38	2016	5
29	51.81	88	26	30	2016	6
30	43.37	89	36	20	2016	7
31	99.23	88	18	34	2016	8
32	71.59	90	18	14	2016	9

[7]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
from sklearn.metrics import mean_absolute_error
```

[8]

```
#Set Variables (X) and Target (Y)
#Divide data into train/test
#Fit the model
X =mikkeller_cleaned.drop(["Sales"], axis=1)
Y = mikkeller_cleaned["Sales"]
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2, random_state = 64)
reg = LazyRegressor(verbose=0, ignore_warnings=False, custom_metric= mean_absolute_error)
models,pred = reg.fit(X_train, X_test, y_train, y_test)
models
```

```
100%|| 41/41 [00:02<00:00, 14.91it/s]
```

LazyPredict Results table is documented in Table 12, LazyPredict results for paper III, predicting Mikkeller sales with Google search data

Appendix 4, python LazyPredict code for paper V, predicting Apple stock price volatility with Google searches

[1]

```
#importing necessary libraries
import pandas as pd
import seaborn as sns
%matplotlib inline
import statistics as st
import matplotlib.pyplot as plt
from numpy import mean
from numpy import std
```

[2]

```
#importing the iphone dataset as a dataframe
apple = pd.read_csv('Apple.csv')
apple.tail()
apple.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 368 entries, 0 to 367
Data columns (total 29 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Week starting             367 non-null    object
 1   Week ending               367 non-null    object
 2   Week no.                  367 non-null    float64
 3   1stLogDiff                366 non-null    float64
 4   Volatility                365 non-null    object
 5   AAPL                      261 non-null    float64
 6   t-1 AAPL                  261 non-null    float64
 7   t-2 AAPL                  261 non-null    float64
 8   t-3 AAPL                  261 non-null    float64
 9   t-4 AAPL                  261 non-null    float64
 10  Open                      261 non-null    float64
 11  High                      261 non-null    float64
 12  Low                       261 non-null    float64
 13  Close                     261 non-null    float64
 14  Avg weekly close          367 non-null    float64
 15  Log(Avg weekly close)     367 non-null    float64
 16  Unnamed: 16               0 non-null      float64
 17  1st diff log(avg close).  366 non-null    float64
 18  Weekly change             351 non-null    object
 19  Adj Close                 261 non-null    float64
 20  Volume                    261 non-null    object
 21  Log(volume)               261 non-null    float64
 22  BlackFriday               277 non-null    float64
 23  Q1 Report                 277 non-null    float64
 24  Q2 Report                 277 non-null    float64
 25  Q3 Report                 277 non-null    float64
 26  Q4 Report                 277 non-null    float64
 27  XmasSales                 277 non-null    float64
 28  iPhoneRelease             277 non-null    float64
dtypes: float64(24), object(5)
memory usage: 83.5+ KB
```

[3]

```
#Clean the dataset as it contains many N/A and we only need 7 columns
#extract relevant columns
apple_clean =apple.loc[:,["1stLogDiff", 'Volatility', "AAPL", "t-1 AAPL", "t-2 AAPL", "t-3 AAPL", "t-4
AAPL"]]
apple_clean
#remove rows with N/A
apple_cleaned = apple_clean.dropna()
apple_cleaned.tail()
#transform volatility to numerical value
apple_cleaned.loc[:, 'Volatility'] = pd.to_numeric(apple_cleaned['Volatility'].str.replace('%', '')) / 100
apple_cleaned.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 257 entries, 107 to 363
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   1stLogDiff  257 non-null    float64
 1   Volatility  257 non-null    float64
 2   AAPL        257 non-null    float64
 3   t-1 AAPL    257 non-null    float64
 4   t-2 AAPL    257 non-null    float64
 5   t-3 AAPL    257 non-null    float64
 6   t-4 AAPL    257 non-null    float64
dtypes: float64(7)
memory usage: 16.1 KB
```

[4]

```
pip install lazypredict
```

[5]

```
from sklearn.model_selection import train_test_split
from lazypredict.Supervised import LazyRegressor
from sklearn.metrics import mean_absolute_error
```

[6]

```
#Set Variables (X) and Target (Y)
#Divide data into train/test
#Fit the model
X =apple_cleaned.drop(["1stLogDiff", 'Volatility'], axis=1)
Y = apple_cleaned["Volatility"]
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2, random_state = 64)
reg = LazyRegressor(verbose=0, ignore_warnings=False, custom_metric=mean_absolute_error)
models,pred = reg.fit(X_train, X_test, y_train, y_test)
models
```

```
100%|| 41/41 [00:04<00:00,  9.17it/s]
```

LazyPredict Results table is documented in Table 13, LazyPredict results for paper V, predicting Apple stock price volatility with Google search data

Paper I: Predicting iPhone Sales from iPhone Tweets

In the Proceedings of the 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, 1–5 September 2014, Ulm, Germany.

Predicting iPhone Sales from iPhoneTweets

Niels Buus Lassen[1], Rene Madsen[1]

[1]Computational Social Science Laboratory

Department of ITM, Copenhagen Business School

Howitzvej 60. 2.14, Frederiksberg, 2000, Denmark

nbl@evalua.dk; Rene.Madsen@infomedia.dk

Ravi Vatrapu[1,2]

[2]Mobile Technology Laboratory

Norwegian School of Information Technology

Schweigaardsgate 14, Oslo, 0185, Norway

vatrapu@cbs.dk

Abstract

Recent research in the field of computational social science have shown how data resulting from the widespread adoption and use of social media channels such as twitter can be used to predict outcomes such as movie revenues, election winners, localized moods, and epidemic outbreaks. Underlying assumptions for this research stream on predictive analytics are that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product and that the shared digital artefact that is persistent can create social influence. In this paper, we demonstrate how social media data

from twitter can be used to predict the sales of iPhones. Based on a conceptual model of social data consisting of social graph (actors, actions, activities, and artefacts) and social text (topics, keywords, pronouns, and sentiments), we develop and evaluate a linear regression model that transforms iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks. This strong correlation between iPhone tweets and iPhone sales becomes marginally stronger after incorporating sentiments of tweets. We discuss the findings and conclude with implications for predictive analytics with big social data.

Keywords: data science, computational social science,social data analytics, predictive analytics, iphone sales, iphone tweets, twitter

I.1. Introduction

Social media has evolved into vital constituents of many human activities. We share aspects of our lives on Facebook, Twitter, Instagram, Tumblr, and many other social media platforms. The resulting social data is persistent, archived, and can be retrieved and analyzed. Social data analytics is not only informing but also transforming existing practices in politics, marketing, investing, product development, entertainment, and news media.

In this paper, we analyze a complex product that generated a large number of opinions on social media. If social media can be characterized as second life for some, then smartphone has evolved into an extension of human body and mind. The product under analytical consideration, Apple iPhone is one of the best-selling products in history and is associated with large amounts of big data on most social media channels. Our paper demonstrates how Twitter social data can be used to predict the future sales of the Apple iPhone. In particular, we analyze the mathematical relationship between twitter social data and iPhone smartphone sales. Our research question is stated below:

Can big social data predict the sales of smartphones?

Our research hypothesis is that smartphone sales are correlated with tweets and can be predicted on the basis of Twitter data. We adopt the method of Asur & Huberman [1] and examine if the same principles for predicting movie revenue with Twitter data can be used to predict iPhone sales. That is, if a tweet can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. We report and discuss a regression model that can predict iPhone sales with 5-10% average error.

The remainder of the paper is organized as follows. Related work on predictive

analytics is reviewed in the next section. Theoretical framework section discusses the AIDA sales funnel model and the Hierarchy of Effects information processing model of advertising. Methodology section discusses twitter data collection and statistical modelling. Results section presents the empirical findings in terms of the regression model. Discussion section offers substantive interpretation of the statistical results and concludes with implications for predictive analytics in particular and computational social science in general.

I.2. Related Work

We deliberately limit the review of extant literature to empirical work that examined the relationship between social data measures (such as facebook posts/likes/comments/shares, and twitter tweets/re-tweets/mentions/polarity etc.) and real-world business outcomes (revenues, stock price etc.).

I.2.1. Social Data & Business Outcomes: Data Science

There has been substantial research work [2-7] in the direction of predicting the stock prices of the companies based on the analysis of content from the online media such as news items, web blogs, twitter feeds. For example, Gavrilov et al., [5] applied data mining techniques on the stock information from various companies by clustering them according to their Standard and Poor (S&P) 500 index, whereas the content from the weblogs is used by Kharratzadeh & Coates [6] to identify the underlying relationships between the companies to make predictions about the evolution of stock prices.

The most notable papers in this regard is from Asur & Huberman [1] showed that social media feeds can be used as effective indicators of the real-world performance. In their work, they used analysis of hourly rate of tweets about movies, their re-tweets and sentiment polarity to accurately forecast the box-office movies revenue. In fact, their prediction of movie revenues based on the social data measures from twitter outperformed the leading market-based predictions of the

Hollywood Stock Exchange. In terms of macro-societal relationships, a research study investigated whether the public mood as measured from large-scale collection of Twitter tweets can be correlated or even predictive of Dow Jones Industrial Average (DJIA) values has been explored by Bollen and Mao [3].

I.2.2. Social Media Analytics: Information Systems

Previous literature about social media analytics have focused upon user-generated content (UCG) [8-10], as well as the organizational [11, 12], business intelligence [13, 14], and predictive aspects of social data [15-20]. For example, Zimbra et al., [10] combined sentiment analysis with topic analysis in order to analyze a Wal-Mart discussion forum to improve organizational decision-making. Huber et al., [8] studied how companies can use wall posts and comments on Facebook to stimulate user engagement, while Lin and Goh [9] investigated the co-existence of customers and marketers in order to determine the value of their content on social media. Heath et al., [11] empirically studied how a strategic organizational engagement in social media can advance organizational goals, while Larson and Watson [12] introduced a social media ecosystem to explain the different stakeholder positions in and around the company. Dinter and Lorenz [13] articulated a research agenda for social business intelligence (social BI), while Rosemann et al., [14] sought to advance the conceptual design of BI with data identified from social networks amongst others through a discussion of social customer relationship management (social CRM) and social BI.

There is an elaborate body of work done on predictive analytics. Seebach et al., [18] suggested that companies include data on customer's online search into their IT systems in order to increase their sensing abilities and create a more agile business. vd Reijden & Koppius [21] studied how online buzz predicts actual sales across different phases of a product lifecycle. Geve and colleagues [15] used Google's index of internet discussion forums and Google's search trends to predict

sales, while Wu and Brynjolfsson [19] used internet searches to predict housing prices. Zhang and Lau [20] developed a business network-based model to analyze and predict business performances (using the proxies of stock prizes). Nann, Krauss, and Schoder [22] analysed multiple online public data platforms such as Twitter and Yahoo! Finance in order to predict the stock market, while Oh and Sheng [17] analysed the predictive power of micro blog sentiments on stock price directional movements.

In general, we find that most of prior related work in the field employs analytical methods for sentiment analysis of the content (social text analytics) or the social network analysis techniques to study social relationships (social graph analytics). When compared to the prior related work, our approach in this paper is novel in the sense that we use both social graph analysis combined with social text analysis (e.g. sentiment analysis) to compute relationship between the social data (e.g. twitter data) and financial performance (e.g. quarterly revenues) of the companies.

Furthermore, as far as we know, we are the first to use twitter data in measuring the relationship between twitter data and quarterly sales of iPhones. That said, we contribute to the knowledge base by empirically investigating a new domain (smartphone sales), theoretically grounding our analysis in relevant domain theories (AIDA & Hierarchy of Effects, discussed next), and extending Asur and Huberman's [1] model to include seasonal weighting.

I.3. Theoretical Framework

In this paper, we build on and substantially extend the method of Asur & Huberman [1] for predicting movie revenue with Twitter data to predict iPhone sales. That is, if a tweet can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or recommend it. In the next section, we discuss the AIDA and Hierarchy of Effects models in order to delineate the

conceptual relationship between users' propensity to tweet and the probability to purchase a product.

I.3.1. AIDA

AIDA model stands for *Awareness, Interest, Desire,* and *Action* and refers to the various stages in a sales process. AIDA was first formulated by Elmo St. Lewis and its original criteria have been subsequently modified to fit technological developments as well as changes in consumer behavior [23]. In terms of the relationship between social data about and sale of an iPhone, the AIDA sales funnel is outlined below.

The first step, *awareness/*attention can result from

- news reading
- friends, colleagues, classmates having the iPhone
- tweets, facebook news, other social media info
- commercials
- seeing the iPhone in use on the metro/bus/train etc

The second step, *interest*/knowledge/liking can result from

- role models having the iPhone
- trying a friend's iPhone
- comparing the iPhone with models from Samsung, Nokia etc. in a mobile phone shop
- reading reviews of phones online including social media

The third step, *desire*/preference can involve

- evaluating iOS vs. Android vs. windows mobile, and forming preferences for what is perceived to be most easy, intuitive, cool, nerdy, configurable, less app costs, most apps etc.

- social influence processes of identification, conformity etc. [24, 25]
- price/needs/features – nice-to-have vs. need-to-have considerations

The fourth and final step, *action*/conviction/purchase, can lead to

- purchase of the new iPhone or one of its competitors.
- holding on to the old mobile/smartphone for a further period
- opting out of the product category of smartphones all together
- product mention/recommendation/review in face-to-face settings (traditional Word of Mouth) and/or online including social media platforms such as twitter

I.3.2. Hierarchy of Effects (HoE)

Hierarchy of Effects (HoE) refers to a family of psychological models that seek to explain human information processing of advertisements [26]. It was first formulated by Lavidge and Steiner [27] and has been the subject of much debate in advertising research [28]. HoE posits a psychological cascade of *cognition*, *affect*, and *behavior* in terms of how advertisements work. According to HoE models, advertisements are processed during the *cognition* phase, leading to the formation of a positive, negative or neutral *affection* which in turn leads to subsequent *behavior*. There are three different orderings of the hierarchy [29]:

- *Learning Hierarchy (C-A-B)* is the typical consumer behavior scenario of learning about a product, forming an opinion, and deciding to purchase it or not.
- *Dissonance Hierarchy (B-A-C)* also known as "buyer's remorse" results when consumers purchase the product first without much deliberation and then have negative experiences of it leading to product awareness.
- *Low-Involvement Hierarchy (B-C-A)* occurs in cases of habitual repurchases owing to brand loyalty (Apple iPhone in our case) and/or product type (for example, bottled water)

Tweets about iPhones can play a role on all three different orderings of the HoE listed above in terms of learning about the product, evaluating one's own experience of it with those of others, and engaging with the product as a brand loyalist by following iPhone related twitter streams. Figure 1, taken from [30], shows the close relationship between the AIDA and HoE models.

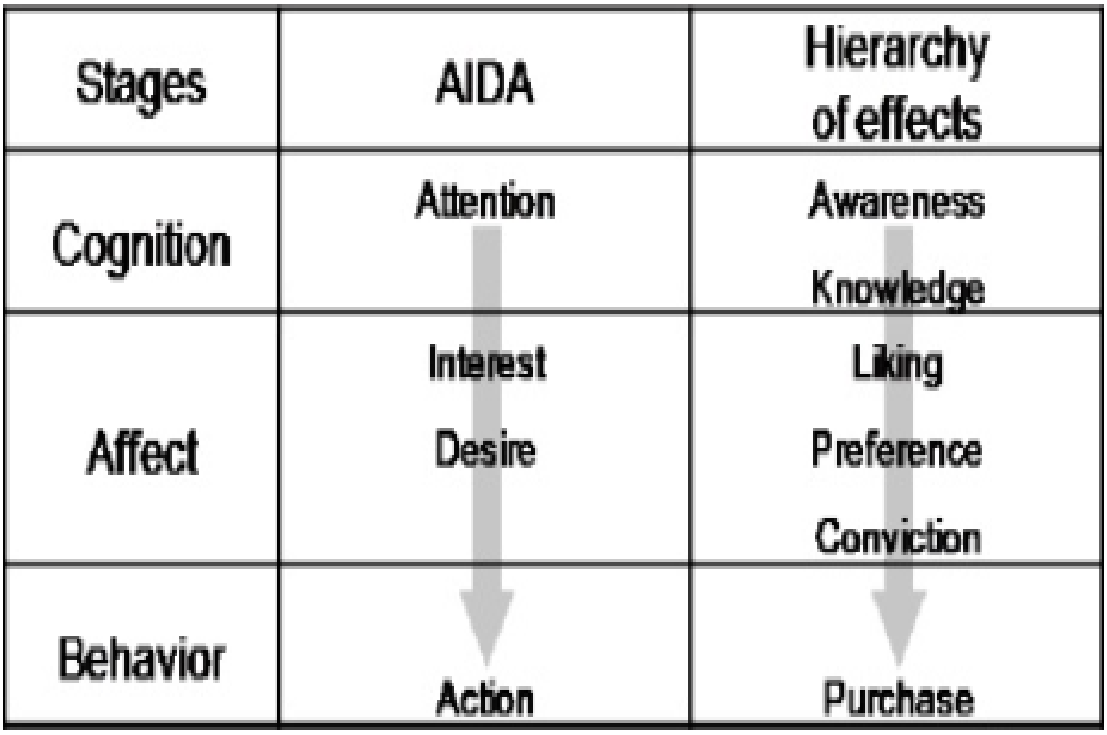


Figure 15: AIDA and Hierarchy of Effects Models

To sum up, tweets about iPhones in particular and smartphones in general are associated with all four stages of the AIDA model and all six stages of the Hierarchy of Effects model. Drawing on Asur and Huberman [1], we treat social data from twitter as a proxy for a user's attention towards the object of analysis which in our case is the iPhone. That said, from the specific domain, we consider a tweet about an iPhone as a proxy for a user's involvement in one of the different stages in the AIDA and HoE models. To be clear, we do not classify each tweet as belonging to a particular stage of AIDA or HoE but treat them as social media

manifestations of real-world activities of users/consumers with respect to the iPhone.

I.4. Methodology

I.4.1. Dataset

We collected over 400 million tweets containing the phrase "iPhone" in the period 2007-2013 using Topsy Pro Analytics[1]. Technically, our data collection did not use the Twitter firehose, but a Twitter API solution with full access to all Twitter data. We searched for the phrase "iPhone" in Topsy Pro, which then returned number of all tweets (Tweets, retweets, and replies) for the time period specified, and with sentiment numbers calculated. These numbers form the basis for prediction of one quarter sales of iPhones.

We read the numbers of Tweets, and corresponding sentiment number in Topsy Pro on the screen, and inputted those numbers into Microsoft Excel. We employed calendar based quarters rather than the financial quarters of Apple for the modeling.

I.4.2. Quantity of Tweets

To provide an example, for the time period of 10-September-2013 to 10-December-2013, we made a data query in Topsy pro, specifying the period and searching for the phrase "iPhone" in all tweets (tweets, replies, retweets). For this example result was 44.62 million tweets and the corresponding sentiment number of 64.

I.4.3. Quality of Tweets

The sentiment number in above example expresses 64% of all tweets as positive. The Topsy Pro has calculated this sentiment number on a smaller fraction of the 44,62 mio tweets. The Topsy Pro sentiment algorithm is a black box, and all we know, from their self-reported descriptions, is that it is optimized for English text.

[1] https://pro.topsy.com/

If we define

- *p : Tweets with positive sentiment*
- *n : Tweets with negative sentiment*
- *o: Tweets with neutral sentiment*
- *t : Total number of Tweets*

then Subjectivity is:

$$Subjectivity = \frac{p+n}{o} = \frac{p+n}{t-(p+n)} \quad (1)$$

and Positivity to Negativity (PN) Ratio is:

$$PNRatio = \frac{p}{n} \quad (2)$$

In Topsy pro the equivalent value is a normalized ratio (0 - 100%) between the positive tweets and tweets with opinions

$$Sentiment = \frac{p}{p+n} \quad (3)$$

I.4.4. Seasonal Weighting of Tweets

Season weight was calculated as the given quarter's proportion of the last calendar year. For example, the season weight for calendar Q3.2013 was calculated as below:

$$\frac{\text{Q3.2013 iPhone sales}}{(\text{Q3.2013} + \text{Q2.2013} + \text{Q1.2013} + \text{Q4.2012})}$$

$$= \frac{\text{33.8 million iPhone sales}}{(33.80 + 31.24 + 37.43 + 47.79)}$$

$$= 0.225$$

This proportion number 0.225 is then divided with 0.25 (0.225 / 0.25 = 0.90) to

yield the season weight for that particular quarter. So the season weight for Q3.2013 is 0.90 which is multiplied with the 38.72 million tweets for that quarter.

Calculating season weights this way, always 4 quarters back in time, ensures that the calculation is always a mix of Q1, Q2, Q3 & Q4. So only one season weight has to be estimated, which is the latest number for prediction for next quarter. We also tried with 2 years average on the season weighting calculation, but best correlation between iPhone tweets and iPhone sales was obtained with calculation of season weight for 1 year of sales data. The season weighting method with best correlation is based on 1 year of sales data, so an estimated season weight must always go 1 year back. It might be critiqued that once the model get the season weight, it gives the model a strong hint on the number of sales. We do not agree this criticism as most sales prediction models incorporates season weights, as sales fluctuates with considerable season variation. Our use is not much different from the use of season weights in other prediction models.

I.4.5. Overall Model

We have made both a linear regression, and a multiple regression prediction model, based on Twitter data. Our final choice was to include the sentiment data from Topsy Pro[2] as our second variable as the sentiment variable improved the correlation and accuracy of the prediction model. Input for the prediction model was then:

$$y = \beta_a \cdot A_{tw} + \beta_p \cdot P_{tw} + \alpha + \varepsilon \qquad (4)$$

Where

- A_{tw}: *Time lagged and season weighted Twitter data*

[2] https://pro.topsy.com

- P_{tw}: *Sentiment of* A_{tw}
- *y: iPhone sales in Units*

After using multiple regression analysis in SAS statistical software, we could calculate difference between predicted sales and actual sales, which ended up with 5-10% average error. This concludes our methodological discussion and we now present and discuss the results.

I.5. Results

As mentioned earlier, we used Topsy Pro, to analyze over 400 million tweets in the period of the Third Quarter of 2007 to the Fourth Quarter of 2013 (Q3.2007-Q4.2013). As Apple publishes iPhone sales by quarters, it became natural to build a prediction model that worked quarterly. A monthly sales prediction model would involve the same principles but our model building followed the structure of quarterly sales data.

Over the period Q3.2007 – Q4.2013 there has been a natural development in the size of the population that is active on Twitter. The development in Twitter users from 2010-2013 could have affected our prediction model. However, from a statistical standpoint, Twitter users showed the same usage patterns during 2010-2013 when tweeting about the iPhone. We did leave out 2007-2009 from our model building for the main reason was a weak link between tweets and sales. 2009 was just atypical in many ways, a statistical outlier – and would have worked as noise for our regression model. From 2010 onwards it is the period of iPhone 3GS, 4, 4S, 5, 5C & 5S.

There is a strong and documented correlation between tweets and iPhone sales in the 2010-2013 period with the Rsquare coefficient of 0.95 and 0.96 for multiple regression with sentiment as the second variable. Output from SAS statistical program is available in the Appendix. Multiple regression analysis – year for year

– is a straightforward and quite easy process. However, modeling on a quarterly basis, is a different matter. Only the introduction of seasonal weighting could make our regression model work on a quarterly basis. We have observed that many other prediction models like Morgan Stanley's "Alphawise Smartphone tracker" also use seasonal weighting. We did not copy the principle of seasonal weighting from others, but based on our practical model building professional experience, we realized the necessity of quarterly seasonal weighting. The principles for monthly weighting, would follow – more or less – the same principles if monthly sales data is available. We ended with a prediction model, which showed an average error on app 5-10% for most of the time periods with iPhone sales. The 5-10% average error is close to the average error of the leading predicting methods from Morgan Stanley and IDC – and our model is much simpler and uses less factors (discussion forthcoming). With more research into our model, we expect to get the average error even further down. Figure 2, below presents predicted vs. actual iPhone sales.

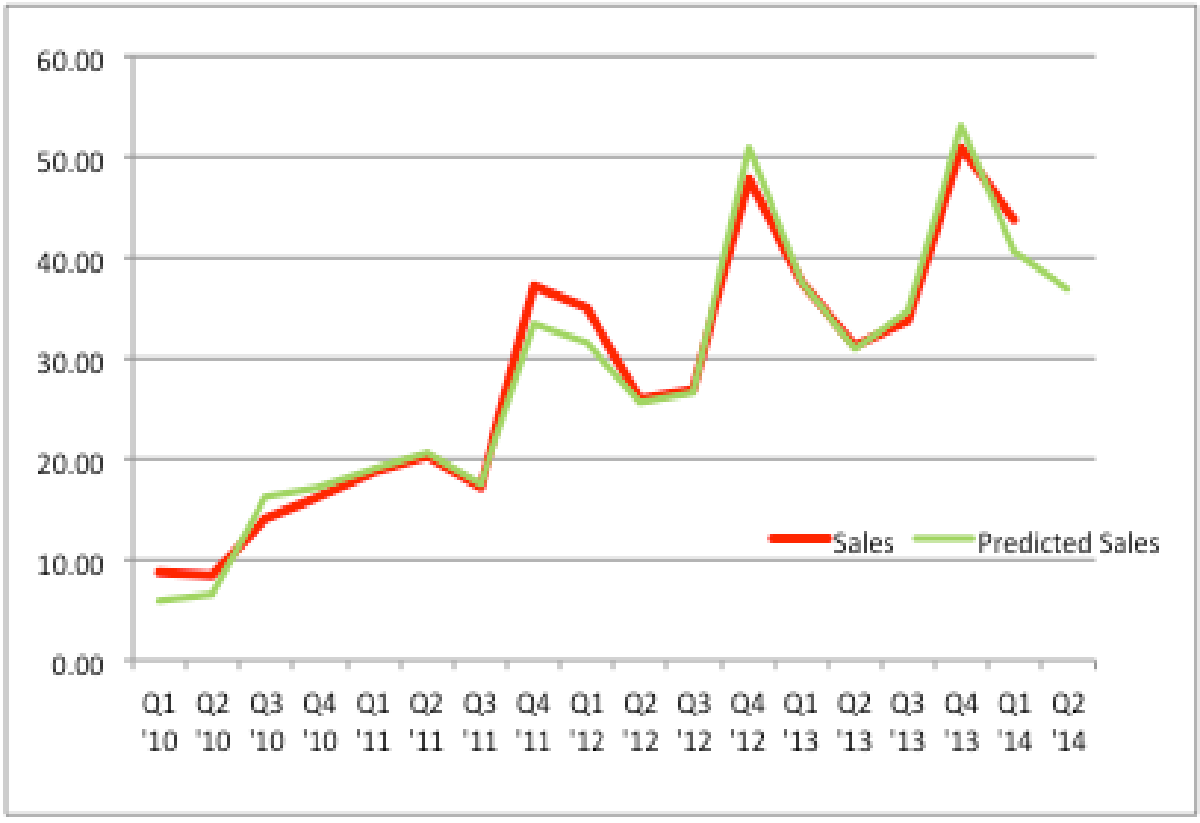


Figure 2: Predicted Quarterly Sales vs. Quarterly Sales

Our main finding is the strength of Twitter as a social data source for predicting smartphone sales. We assume the principles of our prediction model can be used on other products that generate customer opinions and feelings on Twitter. Figure 2 presents the model with final data and shows a prediction of 37 million iPhone sales for Q2.14.

Figure 3 shows that the subjectivity has a declining tendency over time suggesting that people are not as opinionated (passionate) about iPhones as they used to be. This is consistent with the fact that the latest versions of iPhone have not gained any major technological innovations but has shifted from "better" to "more" as in more CPU power, pixel density, and memory. There is a spike in 2011 Q4 around the introduction of iPhone 4S. Also many other black touch sensitive HD screen smart phones with similar capabilities and competitive prices have been introduced on the market since 2010. As the smartphones have increasingly become a mass market product, the "cool" factor of the iPhone has diminished.

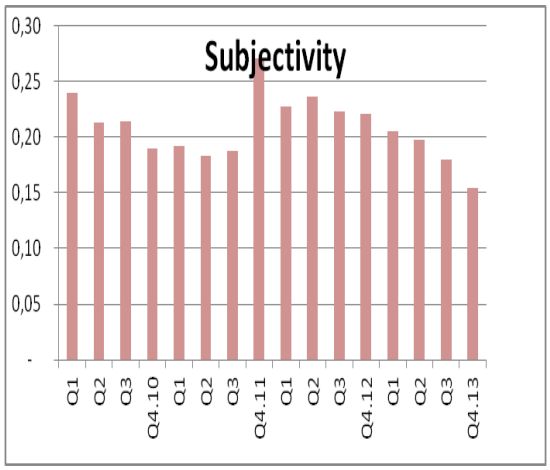


Figure 3: Subjectivity values based on formula (2)

Both the PNRatio shown in Figure 4 and the Sentiment ratio shown in Figure 5 shows a declining tendency that indicates that people are still positive about iPhones but with the overall tendency is decreasing over time. This is consistent with the subjectivity findings as people are less opinionated and less positive about iPhones than before.

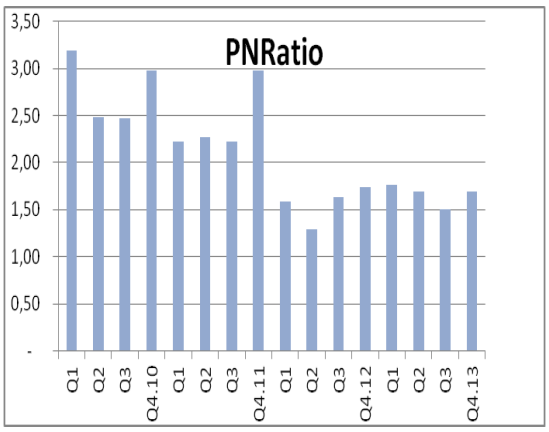


Figure 4: PNRatio values based on formula (3)

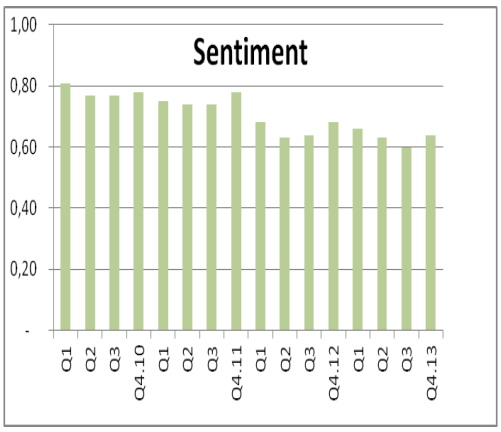


Figure 5: Sentiment values based on formula (4)

Figure 6 presents the output from the statistical software, SAS.

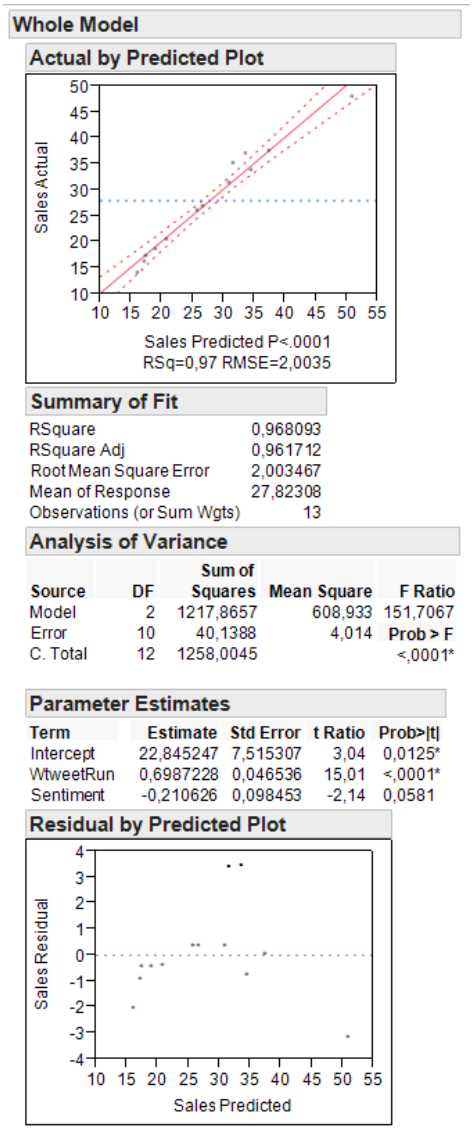


Figure 6: SAS Output for the Prediction Model

I.6. Discussion

To summarize, we used Topsy Pro, to analyze over 400 million tweets in the period of the Third Quarter of 2007 to the Fourth Quarter of 2013 (Q3.2007- Q4.2013). We have made both a linear regression, and a multiple regression prediction model, based on Twitter data. Previous research has explored the differences between tweets, retweets and replies on Twitter [31, 32]. However, for our initial model building, we used all the tweets about the iPhone with no differentiation between tweets, retweets, and replies and also with no sentiment analysis.

We treated all the tweets as equal and built the first model. Trying to model with 1, 2 and 3 of the types of tweets, retweets & replies, it became obvious that modeling on all types of tweets (tweets, retweets & replies) gave the best correlation between twitter activities and iPhone sales. One of the metrics of evaluating the impact of tweets and the engagement of followers is called exposure. The exposure of a tweets is calculated as the total potential impressions it has, that is the sum of all followers including each retweet and the sum of their followers and so on. This gives an estimation of the maximum possible users that had the opportunity to read the tweet. It does not remove overlap in users, is simple to calculate and gives a relative performance count to track twitter trends. As a proxy for attention we have chosen only to count original tweets, retweets and replies since these represent active measurable involvement of users.

Social data (like all data) suffers from seasonal variations and therefore requires a cautious approach to extracting the underlying trend. Likewise with sales, for example, smartphone's are a typical Christmas present and have a boosted sales in Q4. To follow the domain-specific theoretical models of AIDA and HoE models, we considered time lagging Twitter data from the beginning. When building the prediction model, we learned that quarter to quarter correlation between Twitter data and iPhone sales did not have the best correlation. We could improve this correlation substantially by pushing Twitter data back in time. We tried many combinations, with 3-6 months of Twitter data, as basis for quarterly sales. For our model building, we chose to weigh all quarterly Twitter data after season weights. Season weights were calculated as the quarter´s sales proportion of a full year. The quarter up for prediction, calendar Q4.2013, got a season weight as a 2 years average, of the season weights in Q4.2012 & Q4.2011. From Adstock models, and other related sales prediction models based on AIDA, we know there is a timelag from customer attention to the actual product purchase. We therefore tested on

Twitter data, timelagged back in time – in relation to the quarter we tried to predict. We tried many timelags back in time, and ended up with best correlation between iPhone tweets and iphone sales, for Twitter data pushed back 20 days. An example with predicting calendar Q4.2013: Topsy Pro extract of Tweets containing the phrase "Iphone" and belonging sentiment number, for the period 10 sep 2013 - 10 dec 2013 – which is the basis for predicting calendar period Q4.2013 (1 oct. 2013 – 31 dec 2013). So, the prediction model only predicts quarter sales 20 days before the quarter ends. And 50 days before Apple releases the sales figures.

Our final choice for the model-building was to include the sentiment data from Topsy Pro[3] as our second variable as the sentiment variable improved the correlation and accuracy of the prediction model. Regarding the quality of tweets, the sentiment numbers corresponding to given 3-month period of Twitter data was calculated automatically by the sentiment algorithm of Topsy Pro. As such, the sentiment analysis method is a black box. It is described the algorithm is optimized for English text, and for our 400 million tweets, the majority is English text. For the non-English tweets? In practice, the sentiment numbers improved the correlation between iPhone twitter data and iPhone sales. So we conclude that the Topsy Pro sentiment algorithm also works on non-English text, but presumably with a lower accuracy than on English text.

Our final model is then:

$$y = \beta_a \cdot A_{tw} + \beta_p \cdot P_{tw} + \alpha + \varepsilon \quad (4)$$

[3] https://pro.topsy.com

Where

- A_{tw}: *Time lagged and season weighted Twitter data*
- P_{tw}: *Sentiment of* A_{tw}
- *y: iPhone sales in Units*

We model the relationship between iPhone sales and iPhone tweets in the period of 2010-2013 and exclude the period of 2007-2009. We find the data for time period of 2007-2010 to be noisy. But from 2010 – 2013 the statistical association is relatively stable, and gives an excellent correlation. Potential reasons could be historical development of user base on Twitter, and also development of the socio-cultural practices of using twitter. We observed a 5-10% average error from our prediction model in formula (1) with the actual sales data over a 2 year period 2012-2013. This average error is not far from the predictions of Morgan Stanley and IDC. For benchmarking purposes, we have identified a few leading prediction methods.

- Morgan Stanley´s "Alphawise Smartphone tracker" by Katy Huberty based on Google trend data, seasonal weighting, and socio economic factors[4].
- IDC's *Worldwide Quarterly Mobile Phone Tracker®, uses* bottom-up methodology[5]
- Steve Milunovich at UBS[6]
- Peter Misek at Jefferies[7]

Generated By	Information Categories	Typical Investment Debates	Typical Applications
Businesses	Operating Locations	Do a company's operating locations offer a strategic advantage over its competitors	Emerging markets growth; Company competitiveness
	Product Availability	How is the company positioned to meet demand?	Supply-chain bottlenecks; Demand Estimates

[4] http://tech.fortune.cnn.com/tag/alphawise/
[5] http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37
[6] http://www.forbes.com/sites/chuckjones/2013/12/03/ubs-analyst- milun ovich-upgrades-apple-to-buy-with-650-price-target/
[7] http://www.forbes.com/sites/chuckjones/2013/09/13/jefferies-peter-mi sek-says-terrible-yields-on-iphone-fingerprint-sensor-hurting-production/

	Product Pricing	Is the company able to maintain its prices vis-à-vis its competitors?	Company/Sector Margin Pressure; Inflation; Inventory Growth
	Company Hiring	What positions is the company hiring for?	New product expectations; Growth expectations
Consumers & Clients	Demographics	What is the relative demand from different regions?	Performance of new vs. existing stores/regions
	Product Interest	How successful would a new product launch be? Demand trends	New Product demand Sector demand Consumer spending
	Brand Interest	How is a company's market share evolving	Market share changes

Table 1

Source: Morgan Stanley Research, AlphaWise

None of the corporate market research analysts reveal the technical background for their prediction methods. One of the best predictions comes from Alphawise Smartphone tracker and we shortly compare it to Huberman's model [1]. There is nothing public about the math in this model but there is public description in general terms some of the methodology behind the AlphaWise approach[8]. The generic AlphaWise model is very complex as it takes a vast number of factors into consideration. The factors consist of both Business factors such as Location, Availability, Pricing, and Hiring and Customer related aspects such as Demographics, Product Interest and Brand interest as shown in Table 1. Which of the factors are actually included in the Smartphone tracker application is unclear and it is a qualified guess that Morgan Stanley uses multiple regression.

We did not choose to analyze Samsung Galaxy smartphone sales as “Galaxy” is a common phrase and will create problems when analyzing it on Twitter. On the other hand, the iPhone is a unique smartphone name and is one of the most tweeted products. These were the main criteria for our selection of the iPhone, as a case for

[8] http://tinyurl.com/q2bkxcd

a Twitter prediction model. We believe that such technical matters will increasingly become important factors in how companies choose product names. Uniqueness of the product name and hence a possibility for conducting social media analytics will be a point of consideration in the future. This applies for prediction models, customer insights, and many other analytical disciplines that deal with social data.

Regarding generalization, we believe that our approach does generalize to other products of predictions for future years. Different products will require different season weights but building the prediction model of two different products will follow the same principles, with two different set of season weights. The time lag can also be different from product to product. For example, some products could be best predicted with 5 months of Twitter data. Ultimately, the prediction of sales from social data depends on how that specific product´s consumer psychological decision-making process is mirrored on social media channels such as twitter and facebook. Some products will have strong correlation between product posts on social media and product sales in retail and web shops, and some will show weak correlation.

We did consider System Dynamics mathematics, as a model. System dynamics was created during the mid-1950s by Professor Jay Forrester of the Massachusetts Institute of Technology based on a dynamic complex set of differential equations, and causal data relationships. One of the authors of this article have used System Dynamics to predict Christmas tree export from Denmark to Germany. System dynamics is more optimal for complex data pictures containing significant production cycles. It would be possible to build a system dynamics prediction model also containing twitter data, to predict smartphone sales. A System Dynamics prediction model for smartphone sales could be a natural sequel to this article.

We chose not to experiment with Facebook data for our model building, based on the fact that many product pages on Facebook have about 1% user activity – so for the prediction of smartphone sales, we thought that Facebook was too weak a data source. However, emerging research results are reporting strong correlations between quarterly sales and facebook interactions such as posting, commenting, liking, and sharing [33, 34]. That said, for more in-depth analysis of the smartphone sales, one could include big data analysis of social data from Facebook and other leading social media channels such as Tencent in China. A clear advantage of predicting sales with twitter data is the real-time access to data through Topsy Pro and other analytical tools. Changes in trends and the market can be identified with almost no delays. There is no requirement of phone interviews and traditional observations of customer behavior in this social media analytical approach.

I.6.1. Implications for Organizations

Our research results have several direct and indirect implications for organizations. The direct implications, obviously, are that sales can be predicted from social media datasets. The indirect implications are that organizations should strategically engage, analyze, and manage social media platforms and mobile applications given the strong correlations between real-world sales and digital-world activities such as social media interactions. An informed and intelligent organizational use of social media to generate competitive advantages [35] requires not only a the adoption of use of technological artefacts for creating valuable affordances [36] for users/consumers but also an understanding of the psychological aspects of how and why consumers share their experiences, interactions, and opinions about products and services as facebook posts, Instagram pictures and tweets [37].

As stated earlier, we believe that the principles of our prediction model can be used on other products that generate customer opinions and feelings on Twitter. In our opinion, big social data analytics that is informed by domain-specific models and

theories such as the AIDA (Attention, Interest, Desire, and Action) and the HoE (Heirarchy of Effects) models can yield descriptive, prescriptive, and predictive insights. On that note, we think that the novelty and contribution of our work is in the fact that we conduct theory based big social data analytics (in our case, marketing theories of AIDA and HoE). We believe that this is a small but substantial step towards generating causal explanations and not being limited to documenting statistically significant correlations of sales and social media interactions.

I.7. Conclusion

Drawing from the theoretical framework of AIDA and Hierarchy of Effects models in advertising combined with an assumptions that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product, we demonstrated how social media data from twitter can be used to predict the sales of iPhones. We developed and evaluated a linear regression model that transforms iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks. This strong correlation between iPhone tweets and iPhone sales becomes marginally stronger after incorporating sentiments of tweets. We discuss our results in terms of a leading industry research as well as academic research based predictive models.

I. References

[1] Asur, S., and Huberman, B.A.: 'Predicting the future with social media', in Editor (Ed.)^(Eds.): 'Book Predicting the future with social media' (IEEE, 2010, edn.), pp. 492-499

[2] Bakshy, E., Simmons, M.P., Huffaker, D., Teng, C., and Adamic, L.: 'The social dynamics of economic activity in a virtual world', ICWSM2010. http://misc.si.umich.edu/publications/18, 2010

[3] Bollen, J., and Mao, H.: 'Twitter mood as a stock market predictor', Computer, 2011, pp. 91-94

[4] Dorr, D.H., and Denton, A.M.: 'Establishing relationships among patterns in stock market data', Data & Knowledge Engineering, 2009, 68, (3), pp. 318-337

[5] Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R.: 'Mining the stock market (extended abstract): which measure is best?', in Editor (Ed.)^(Eds.): 'Book Mining the stock market (extended abstract): which measure is best?' (ACM, 2000, edn.), pp. 487-496

[6] Kharratzadeh, M., and Coates, M.: 'Weblog Analysis for Predicting Correlations in Stock Price Evolutions', in Editor (Ed.)^(Eds.): 'Book Weblog Analysis for Predicting Correlations in Stock Price Evolutions' (2012, edn.), pp.

[7] Mittermayer, M.-A.: 'Forecasting intraday stock price trends with text mining techniques', in Editor (Ed.)^(Eds.): 'Book Forecasting intraday stock price trends with text mining techniques' (IEEE, 2004, edn.), pp. 10 pp.

[8] Huber, J., Landherr, A., Probst, F., and Reisser, C.: 'Stimulating User Activity On Company Fan Pages In Online Social Networks', ECIS 2012 Proceedings. Paper 188. http://aisel.aisnet.org/ecis2012/188, 2012

[9] Lin, Z., and Goh, K.Y.: 'Measuring the Business Value of Online Social Media Content for Marketers', ICIS 2011 Proceedings. Paper 16. http://aisel.aisnet.org/icis2011/proceedings/knowledge/16 2011

[10] Zimbra, D., Fu, T., and Li, X.: 'Assessing public opinions through Web 2.0: a case study on Wal-Mart', ICIS 2009 Proceedings. Paper 67. http://aisel.aisnet.org/icis2009/67, 2009

[11] Heath, D., Singh, R., Ganesh, J., and Kroll-Smith, S.: 'Exploring Strategic Organizational Engagement in Social Media: A Revelatory Case', ICIS 2013 Proceedings. http://aisel.aisnet.org/icis2013/proceedings/EBusiness/13/, 2013

[12] Larson, K., and Watson, R.T.: 'The value of social media: toward measuring social media strategies', in Editor (Ed.)^(Eds.): 'Book The value of social media: toward measuring social media strategies' (2011, edn.), pp.

[13] Dinter, B., and Lorenz, A.: 'Social Business Intelligence: a Literature Review and Research Agenda', in Editor (Ed.)^(Eds.): 'Book Social Business Intelligence: a Literature Review and Research Agenda' (2012, edn.), pp.

[14] Rosemann, M., Eggert, M., Voigt, M., and Beverungen, D.: 'Leveraging social network data for analytical CRM strategies: the introduction of social BI', in Editor (Ed.)^(Eds.): 'Book Leveraging social network data for analytical CRM strategies: the introduction of social BI' (AIS Electronic Library (AISeL), 2012, edn.), pp.

[15] Geva, T., Oestreicher-Singer, G., Efron, N., and Shimshoni, Y.: 'Do Customers Speak Their Minds? Using Forums and Search for Predicting Sales', Available at SSRN: http://ssrn.com/abstract=2294609 or http://dx.doi.org/10.2139/ssrn.2294609, 2013

[16] Koppius, O.: 'The Value of Online Product Buzz in Sales Forecasting', ICIS 2010 Proceedings. Paper 171. http://aisel.aisnet.org/icis2010_submissions/171, 2010

[17] Oh, C., and Sheng, O.: 'Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement', in

Editor (Ed.)^(Eds.): 'Book Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement' (2011, edn.), pp.

[18] Seebach, C., Pahlke, I., and Beck, R.: 'Tracking the Digital Footprints of Customers: How Firms can Improve Their Sensing Abilities to Achieve Business Agility', ECIS 2011 Proceedings, 2011

[19] Wu, L., and Brynjolfsson, E.: 'The future of prediction: how Google searches foreshadow housing prices and quantities', ICIS 2009 Proceedings. Paper 147. http://aisel.aisnet.org/icis2009/147, 2009

[20] Zhang, W., and Lau, R.: 'The Design of a Network-Based Model for Business Performance Prediction', ICIS 2013 Proceedings. http://aisel.aisnet.org/icis2013/proceedings/KnowledgeManagement/10/, 2013

[21] vd Reijden, P., and Koppius, O.R.: 'The Value of Online Product Buzz in Sales Forecasting', in Editor (Ed.)^(Eds.): 'Book The Value of Online Product Buzz in Sales Forecasting' (2010, edn.), pp. 171

[22] Nann, S., Krauss, J., and Schoder, D.: 'Predictive Analytics On Public Data-The Case Of Stock Markets', roceedings of the 21st European Conference on Information System. http://www.staff.science.uu.nl/~Vlaan107/ecis/files/ECIS2013-0615-paper.pdf, 2013

[23] Li, H., and Leckenby, J.: 'Examining the Effectiveness of Internet Advertising Formats', in Schumann, D., and Thorson, E. (Eds.): 'Internet

Advertising: Theory and Research' (Lawrence Erlbaum Associates, 2007), pp. 203-224

[24] Kelman, H.C.: 'Compliance, identification, and internalization: Three processes of attitude change', The Journal of conflict resolution, 1958, 2, (1), pp. 51-60

[25] Cialdini, R., and Goldstein, N.: 'Social influence: Compliance and conformity', Annual Review of Psychology, 2004, 55, (1), pp. 591-621

[26] Schumann, D., and Thorson, E.: 'Internet Advertising: Theory and Research' (Lawrence Erlbaum Associates, 2007. 2007)

[27] Lavidge, R.J., and Steiner, G.A.: 'A model for predictive measurements of advertising effectiveness', The Journal of Marketing, 1961, pp. 59-62

[28] Barry, T.E.: 'The development of the hierarchy of effects: An historical perspective', Current issues and Research in Advertising, 1987, 10, (1-2), pp. 251-295

[29] Ray, M.: 'Marketing communication and the hierarchy of effects', New models for communication research, 1973, pp. 146-175

[30] Belch, G.E., Belch, M.A., Kerr, G.F., and Powell, I.: 'Advertising and promotion: An integrated marketing communications perspective' (Mcgraw-Hill, 2008. 2008)

[31] Romero, D., Galuba, W., Asur, S., and Huberman, B.: 'Influence and passivity in social media', Machine Learning and Knowledge Discovery in Databases, 2011, pp. 18-33

[32] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P.: 'Measuring user influence in twitter: The million follower fallacy', ICWSM, 10, 2010, pp. 10-17

[33] Mukkamala, R., Hussain, A., and Vatrapu, R.: 'Towards a Formal Model of Social Data', IT University Technical Report Series, 2013, TR-2013-169, pp. https://pure.itu.dk/ws/files/54477234/ITU_TR_54472013_54477169.pdf

[34] Mukkamala, R., Hussain, A., and Vatrapu, R.: 'Towards a Set Theoretical Approach to Big Data Analytics', Proceedings of IEEE Big Data 2014, Anchorage, USA, in press/2014

[35] Vatrapu, R.: 'Understanding Social Business', in Akhilesh, K.B. (Ed.): 'Emerging Dimensions of Technology Management' (Springer, 2013), pp. 147-158

[36] Vatrapu, R.: 'Explaining culture: an outline of a theory of socio-technical interactions', Proceedings of the 3rd ACM International Conference on Intercultural Collaboration (ICIC 2010), 2010, pp. 111-120

[37] Kunst, K., and Vatrapu, R.: 'Towards A Theory Of Socially Shared Consumption: Literature Review, Taxonomy And Research Agenda', Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, 2014, pp. ISBN 978-970-9915567-9915560-9915560

Paper II: Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data

In the Proceedings of the 38th Symposium i Anvendt Statistik, Copenhagen Business School, Frederiksberg, Denmark, 25–27 January 2016.

Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data

Niels Buus Lassen, Ravi Vatrapu, Lisbeth la Cour, René Madsen, Abid Hussain

II.1. Introduction

In this chapter, we will advance a theory of social data that distinguishes between constituent dimensions of social graph (i.e., socio-technical affordances of social media networks) and those of social text (i.e., communicative and linguistic properties of social media interactions) as distinct but complementary elements of predictive big social data analytics. Additionally, to illustrate the validity and applicability of our proposed theory, we adhered to the schematic steps advocated by Shmueli and Koppius (2011) in building empirical predictive models that blend social graph analysis with social text analysis to: (1) compute correlations between social data from multiple social media platforms (i.e., Facebook and Twitter) and the financial performance (i.e., quarterly revenues) of corporate entities (i.e., iPhones and H&M), as well as; (2) make predictions about the future performance of these corporate entities. In doing so, we endeavor to provide an answer to the following research question: *How can big social data analytics be utilized to predict business performance?*

This paper comprises four sections, inclusive of this introduction.. In Section 2, we construct our theory of social data by extending Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of

social media. Section 3 outlines our methodological strategy for extracting and analyzing big social data to build empirical predictive models of business performance. Results from analyzing these empirical predictive models are also reported in Section 3. The last section, Section 4, summarizes the: (1) implications of this study to both theory and practice; (2) insights to be gleaned towards informing the application of predictive analytics to big social data; (3) possible limitations in the interpretation of our empirical findings, and; (4) probable avenues for future research.

II.2. Towards a theory of social data

To bridge the knowledge gaps in extent literature, we advance a theory of social data that extends Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Social media (e.g., Facebook and Twitter), at the highest level of abstraction, involve social entities interacting with: (a) technologies (e.g., an individual using the Facebook app on his/her smartphone), and; (b) other social entities (e.g., the same individual liking a picture of a friend on the Facebook app). Vatrapu (2008, 2010) labelled these interactions as *sociotechnical interactions* (see also Vatrapu and Suthers 2010). Socio-technical interactions yield electronic trace data that we termed as ***social data***. To derive a theory for social data, we must first determine the constituents of socio-technical interactions. As acknowledged by Vatrapu (2010), socio-technical interactions are realized through: (a) a social entity's perception and appropriation of *socio-technical affordances*, as well as; (b) the structures and functions of *technological intersubjectivity* (Vatrapu 2010).

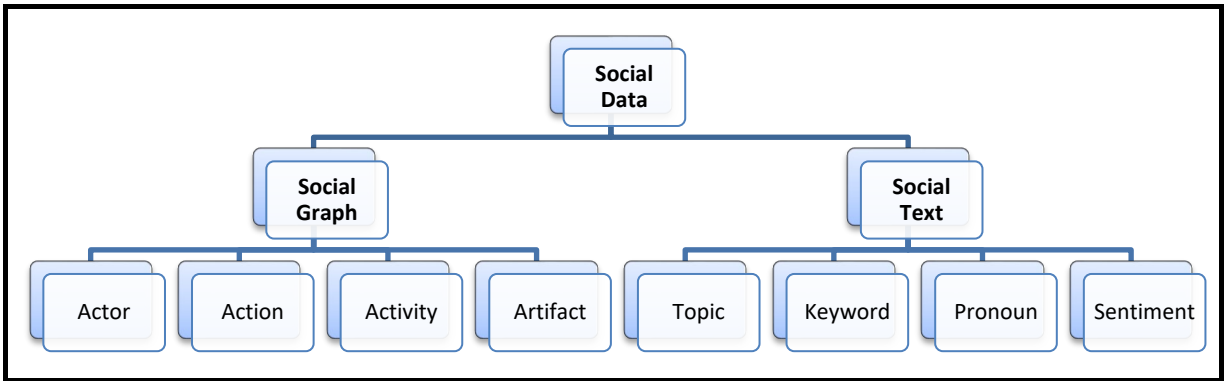


Figure 1: Theory of Social Data

As an illustration of our theory, consider the earlier example of an individual liking a friend's picture on the Facebook app. The performance of such a simple sociotechnical interaction already activates multiple social data elements: an *actor* (i.e., individual) performing an *action* (i.e., liking) on an *artifact* (i.e., Facebook app) for the purpose of expressing a *sentiment* (i.e., like) and contributing to a collective *activity* (i.e., expanding the social network timeline). Such micro social-technical interactions, when amassed in large volumes, constitute the macro world of big social data, the core premise of this paper.

II.3. Methodology and analytical findings

In this section, we presents details about the collection, preparation, exploration, selection, modelling and reporting of two big social data sets to illustrate different aspects of our proposed theory of social data. In general, we adhered to the methodological schematic recommended by Shmueli and Koppius (2011, p. 563) for building empirical predictive models. The remainder of this section is organized in accordance with Shmueli and Koppius's (2011) eight methodological steps of predictive model building as depicted in Figure 2.

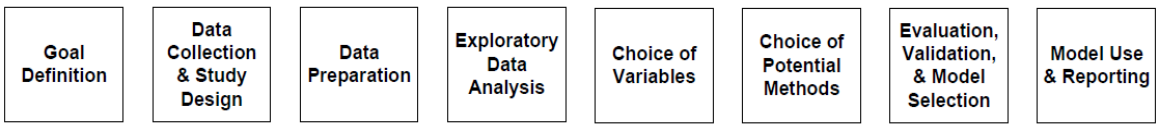


Figure 2: Methodological Steps in Predictive Model Building [Shmueli and Koppius 2011, p. 563]

Step 1: Goal Definition

Our primary goal was to build empirical predictive models of sales from big social data. More specifically, by applying predictive analytics to big social data, we strive to model and accurately predict the real-world numerical outcomes of quarterly sales of Apple iPhone & H&M revenues.

Step 2: Data Collection and Study Design

We discuss the rationale for the study design first followed by details on data collection.

Study Design: The study was designed to collect and analyze big social data sets that serve as illustrative case studies for predictive analytics. Therefore, we deliberately introduce variance into both the predicted variable of sales as well as the predictor variables of social data attributes.

With regard to the predicted variable of sales, we sought to incorporate variance in terms of *product types* (i.e., Apple iPhone: consumer electronics and H&M: fashion-clothes) and *sales channels* (i.e., offline and online; direct and retail). As for the predictor variables of big social data, we incorporated variance in terms of *social media platforms* (i.e., Facebook and Twitter), *theory of social data attributes* (Social Graph: actors, actions, artifacts and Social Text: keywords and sentiment), *dataset sizes* (few millions to hundreds of millions of data points), and *data time periods* (few months to years). Table 1 summarizes the characteristics of the two big social datasets that have been collected, processed and analyzed in this paper.

Table 15: Big Social Datasets Collected for Predictive Analytics

Company	Data Source	Time Period	Size of Dataset	Mapping to Social Data Attributes
Apple[9]	Twitter	2007 → June, 2015	500 million+ tweets containing "iPhone"	▪ **Social Text**: Keyword ("iPhone") ▪ **Social Text**: Sentiment

[9] URL: https://www.apple.com.

H & M[10]	Facebook	January 01, 2009 → March, 2015	~15 million Facebook events	▪ **Social Graph**: Actions (Total Likes) ▪ **Social Graph**: Artifacts (Posts and Comments) ▪ **Social Graph**: Actors (H&M + Non-H&M) ▪ **Social Text**: Sentiment

Data Collection: We now present details on the methods and tools used for data collection for the two big social datasets.

Twitter (Apple: "iPhone")

We collected over 500 million tweets containing the phrase "iPhone" in the period 2007-2015 (till March, 2015) via Topsy Pro Analytics[11]. Technically, our data collection did not connect to the Twitter firehose, but rely on a Twitter API solution with full access to all Twitter data.

Facebook (H&M)

Facebook wall data was extracted by a specialized big social data analytics tool called SODATO. SODATO[12] is an IT artifact, a software solution that is custom built for collecting, storing, processing, and analyzing big social data from social media platforms. The construction of SODATO is not only informed by our proposed theory of social data, but it is also methodologically built in adherence to Sein et al.'s (2011) Action Design Research (ADR) principles. Technically, SODATO utilizes the APIs provided by the social network vendors (e.g., Facebook open source API named as Graph API). Table 2 gives an overview of the social data collected by SODATO from the official Facebook walls of H&M.

[10] URL: http://www.hm.com.
[11] URL: https://pro.topsy.com.
[12] URL: http://cssl.cbs.dk/software/sodato.

Table 16: Overview of Facebook Data					
Company	**Official Facebook Wall: Name (id)**	**Time Period**	**Facebook Posts**	**Facebook Comments**	**Facebook Likes**
H&M	Hm (21415640912)	January, 2009 → March, 2015	127,920	366,863	14,367,067

Sales (Apple and H&M)

Data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. This concludes the presentation of the methods and tools used for data collection and overviews of the different big social datasets. We now discuss the third step in predictive analytics prescribed by Shmueli and Koppius (2011), data preparation.

Step 3: Data Preparation

Twitter (Apple: "iPhone")

We searched for the keyword "iPhone" in Topsy Pro, which then returned number of all tweets (i.e., Tweets, retweets, and replies) for the time period specified, and with sentiment numbers pre-calculated. These numbers form the basis for our prediction of one quarter sales of iPhones. We read the numbers of Tweets, and corresponding sentiment number in Topsy Pro on the screen, and inputted those numbers into Microsoft Excel. We employed calendar based quarters rather than the financial quarters of Apple for the modelling.

Facebook (H&M)

Facebook data was first fetched by SODATO via the Facebook Graph API and was then pre-processed and aggregated in order to make it available on demand for Analytics engine and at the end to the visualization module. The grouping of different analysis units was done in accordance with the different attributes of the theory of social data (Social Graph: actors, actions and artefacts and Social Text: sentiment).

Sales (Apple and H&M)

As mentioned earlier, data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. These were tabulated into Excel spreadsheets together with quarterly measures of social graph and social text.

Step 4: Exploratory Data Analysis

Shmueli and Koppius (2011) stated that during exploratory data analysis: "each question, rather than each construct, would be treated as an individual predictor. In addition to exploring each variable, examining the correlation table between BI and all of the predictors would help identify strong predictor candidates and information overlap between predictors (candidates for dimension reduction)" (p. 657).

Our objectives for the explorative data analytics were twofold: First, to build on the seminal regression model of Asur and Huberman (2010) for predicting movie revenues from twitter sales. Second, based on the Hierarchy of Effects (HoE) (Lavidge and Steiner 1961) and the AIDA (Attention, Desire, Interest, and Action)

(Li and Leckenby 2007) domain-specific models of advertising and sales respectively, to explore different predictor variables, different data transformations of the predictor variables in terms of time lagging and different options for seasonal weighting of the predicted variable, sales.

We organize this section in the order of the two datasets (Apple iPhone tweets & H&M facebook) and describe the explorative data analysis conducted on the respective big social data sets that had already been collected and prepared.

"iPhone" Dataset

For the iPhone dataset, we selected the social data attributes of *social graph: actions* (tweets, re-tweets, replies and mentions) and *social text: keyword* ("iphone") and *social text: sentiments* (positive, negative, and neutral). We explored the temporal dynamics of the social data measures *social graph: actions* and *social text: sentiments* for the filter social text: keyword ("iPhone") directly on the Topsy Pro web site. We then explored the dataset by creating two predictor variables: quantity of tweets and quality of tweets as described below.

Quantity of Tweets

To provide an example, for the time period of September 10, 2013 to December 10, 2013, we made a data query in Topsy pro, specifying the period and searching for the phrase "iPhone" in all tweets (tweets, replies, retweets). For this example result was 44.62 million tweets and the corresponding sentiment number of 64.

Time Lagging of Tweets

As mentioned earlier, our predictive analytics method is informed by both the theory of social data and the AIDA and HoE domain-specific models. The key analytical challenge in social data predictive analytics is to model real-world outcomes from social data measures of social graph (actions, artefacts, activities and actions) and social text (topics, keywords, pronouns and sentiments). From the AIDA and the HoE domain-specific models and based on standard industry practice, we explored different options for time-lagging of social data measures as proxy for the sales funnel inherent in the time period between a potential customer becoming aware of the product, developing an interest in the product, having a desire for it and ultimately deciding to obtain it typically by a sales transaction. We experimented with different time-lags and found 20 days to be the statistically optimal value for the iPhone twitter dataset. As will be discussed later, we found different time lags for different datasets. It is important to note that even though the AIDA and HoE models can help in the exploration of the time lag in the first place and a partial explanation of its existence, they do not theoretically predict a particular value. This, we hope will be addressed with research advances in computational social science in general and predictive analytics in particular.

Seasonal Weighting of Sales

Again, based on the AIDA and HoE models, and given the product life cycle of new models and new operating system releases of Apple iPhone, we conducted season weighting of the quarterly sales. Seasonal weights were calculated as the given quarter's proportion of the last calendar year. For example, the season weight for calendar Q3.2013 was calculated as below:

$$\frac{\text{Q3.2013 iPhone Sales}}{(\text{Q3.2013} + \text{Q2.2013} + \text{Q1.2013} + \text{Q4.2012})} = \frac{\text{33.8 million iPhone Sales}}{(33.80 + 31.24 + 37.43 + 47.79)} = \underline{0.225}$$

This proportion number 0.225 is then divided with 0.25 (0.225 / 0.25 = 0.90) to yield the season weight for that particular quarter. So the season weight for Q3.2013 is 0.90 which is multiplied with the 38.72 million tweets for that quarter.

Calculating season weights this way, always 4 quarters back in time, ensures that the calculation is always a mix of Q1, Q2, Q3 & Q4. So only one season weight has to be estimated, which is the latest number for prediction for next quarter. An estimated season weight for prediction must always go 1 year back. Next, we present the exploratory data analysis of the H&M dataset.

H&M Dataset: Following Shmueli and Koppius (2011)'s advice for exploratory data analysis step of predictive analytics, we explored the predictive power of several different variables constructed from the theory of social data. In summary, we created two categories of the social data attribute of *social graph: a*ctors (H&M and Non-H&M). We then calculated the distribution of the social data attribute of *social graph: artefacts* (posts, comments, and likes) across the two actor types. With respect to the social data attribute of social text: sentiments (positive, negative, and neutral), based on the sentiment analysis of the social text artefacts (posts and comments) discussed earlier, we calculated distributions of sentiments across different kinds of artifacts and actors (i.e. positive sentiments on posts by H&M actors (wall administrators), positive sentiments on posts by Non-H&M actors etc.). We then calculated the quarterly aggregates of these different measures of social data attributes and evaluated the statistical correlation with respect to quarterly sales. Surprisingly, statistically significant positive correlations with quarterly revenues were observed for negative sentiments on total posts.

Logarithmic Transformation and Time lagging of Facebook Likes

Informed by the correlational analysis above and based on further exploratory data analysis with different predictor variables, we selected the logarithmic transformation of 40 days' time lagged total likes per quarter as the main predictor variable from the array of social data attributes listed in Tables 4 and 5 above.

Seasonal Weighting of Quarterly Sales

As with iPhone quarterly sales, we used a weighted measure of the quarterly revenues of H&M to account for seasonal variation of sales corresponding to fashion cycles (i.e., Fall, Winter, Spring and Summer Collections) and holidays across the different H&M markets.

Step 5: Choice of Variables

Choice of the predictor variables is based on careful considerations of theory, domain-specific knowledge and empirical association with predicted variables (Shmueli and Koppius, 2011). Based on exploratory data analysis, the following variables were chosen for the two big social datasets as summarized in Table 3.

Table 3: List of Chosen Predictor Variables

Company / Product	Time Period of Quarter	Seasonal Weighting of Dependent Variable [Sales]	Independent Variable #1 (including info on transformation)	Independent Variable #2	Time-Lagging of Independent Variable #1	Time-Lagging of Independent Variable #2
iPhone Sales (Quarterly)	Calendar Quarters	+	No of tweets over 3 months period	sentiment	20 days	20 days
H&M	Quarter ends 1 month before calendar quarter: Q4.2014 is	+	LOG (No of total likes over 3 months period)	none	40 days	none

	from September 01 → November 30					

Step 6: Choice of Methods

As discussed earlier, we analytical objective was to not only build on but also extend the predictive modelling of Asur and Huberman (2010). As such, we chose regression modelling as the method and sought to extend the method by using time lagged and transformed predictor variables of social data measures and seasonally adjusted predicted variables.

Step 7: Evaluation, Validation and Model Selection

Our overall predictive analytics model for big social data analytics is stated below:

$$y = \beta_a \times A_t + \beta_p \times P_t + \beta_d \times D + \varepsilon$$

Where:

$A_t = \sum A_{st}$
A_{st} = Social media activity in terms of actions by actors on artifacts associated with sales at time t (Social Graph Attributes)
A_t = Accumulated time-lagged social media activity associated with sales at time t
P_t = Polarity at time t (Social Text Attribute)
D = Distribution factor (Sales Channel Attribute)

We now present the specific prediction models for the two big social datasets of iPhone & H&M.

Social Data Predictive Analytics Model for iPhone Sales

We modelled the relationship between iPhone sales and iPhone tweets in the period of 2010-2014 and excluded the period of 2007–2009. While the data for time period of 2007–2009 is noisy, the statistical association is relatively stable for 2010–2013 and gives an excellent correlation. Potential reasons could be historical growth of user base on Twitter, and also the development of socio-cultural practices of using twitter. The predictive model for iPhone sales is:

> Predicted Sales of iPhones Sold (in millions) = WtweetRun * 0,6987228 + Sentiment * (-0,210626) + 22,845247 (intercept)

where

- WtweetRun is the season weighted tweets count for 3 months period time lagged by 20 days back from the sales quarter
- Sentiment is the sentiment for tweets for 3 month period time lagged by 20 days back the from sales quarter

Figure 3 presents the statistical output for the iPhone predictive model.

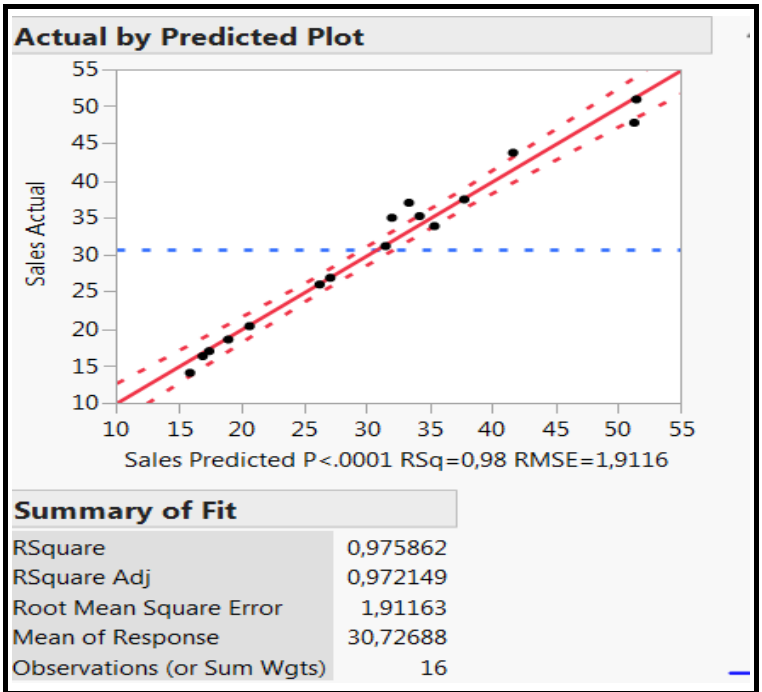


Figure 3: Actual by Predicted Plot of the iPhone Predictive Model

Figure 4 depicts the graph for the iPhone predictive model.

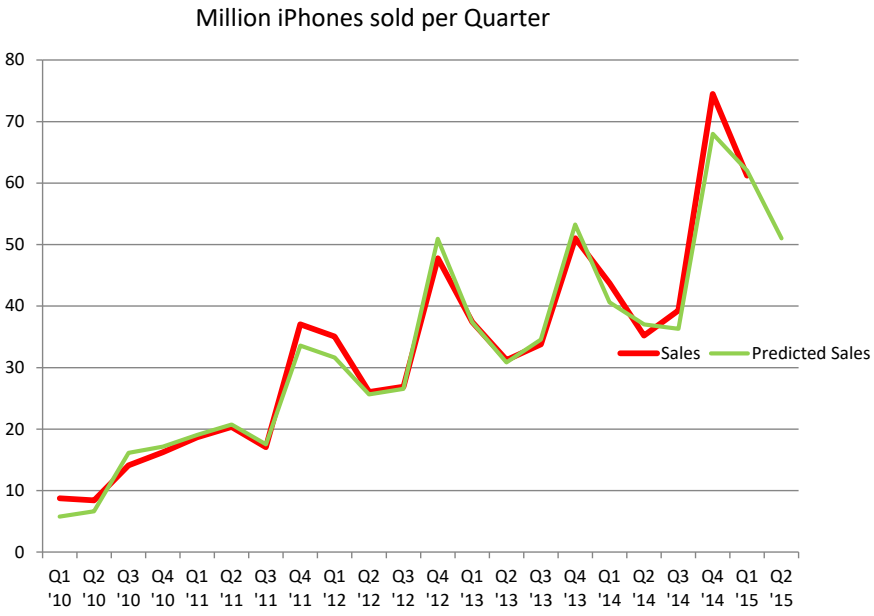


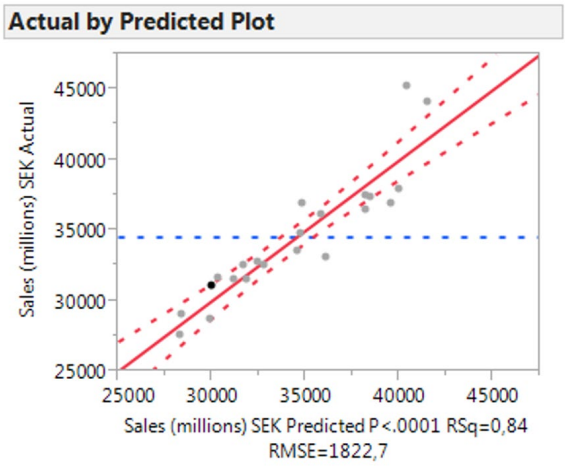
Figure 4: Predicted vs, Actual Sales of iPhones

Social Data Predictive Analytics Model for H&M Revenues

Based on the linear regression for the period 2011-2014, our predictive analytics model for 2014 is given by the following equation:

Predicted Revenue for H&M (in billions SEK) = 2,28 billion SEK * seasonweight * LOG (Facebook total likes time lagged by 40 days back over a 3 months period) + 5,45 billion SEK (the intercept)

Figure 5 presents the SAS output for the 2011-2014 predictive modelling



Summary of Fit

RSquare	0,841863
RSquare Adj	0,833956
Root Mean Square Error	1822,675
Mean of Response	34508,05
Observations (or Sum Wgts)	22

Figure 5: Actual by Predicted Plot of the H&M Revenues [2011-2014]

However, for the period of 2010-2013, based on the linear regression of data for 2009-2013, the predictive model is:

Predicted Revenue for H&M (in billions) = 1,67 billion SEK * seasonweight * LOG (facebook total likes time lagged by 40 days back over a 3 months period) + 13 billion SEK (the intercept)

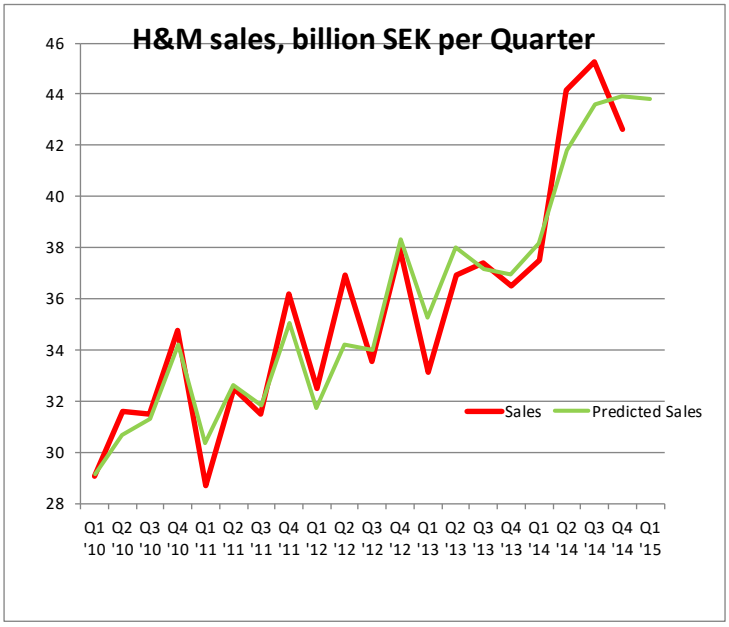


Figure 6: Predicted vs. Actual Revenues of H&M

Figure 6 depicts the combined chart of predicted vs. actual revenues of H&M

Step 8: Model Use and Reporting

In this step, we focus on predictive accuracy and meaning (Shmueli & Koppius, 2011). With regard to our prediction models, we observed a 5-10% average error from our prediction model with the actual sales data over 3 year period 2012-2014. In the case of iPhone, this average error is not far from the predictions of Morgan Stanley and IDC. For benchmarking purposes, we have identified a few leading prediction methods for iPhone sales.

- Morgan Stanley´s "Alphawise Smartphone tracker" by Katy Huberty based on Google trend data, seasonal weighting, and socio economic factors[13].
- IDC's *Worldwide Quarterly Mobile Phone Tracker®, uses* bottom-up methodology[14]

[13] URL: http://www.forbes.com/sites/chuckjones/2014/03/19/morgan-stanleys-alphawise-smartphone-tracker-has-iphone-demand-ahead-of-consensus.
[14] URL: http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37.

- Steve Milunovich at UBS[15]

II.4. Discussion

Though predictive analytics has been touted to be a major growth segment for research into social media, there is only a handful of studies to-date that have managed to capitalize on this opportunity. This paper thus takes a small but concrete step towards furthering this research agenda by advancing and validating a theory of social data for enhancing predictive analytics. Detailed implications for theory and practice are elaborated below.

II.4.1. Implications for Theory

This paper makes a novel contribution to extant literature on several fronts. First, past studies on social networks have typically progressed as two separate research streams with one seeking to comprehend the structural properties of such networks (i.e., social network analysis) (e.g., Johnson et al. 2014; Moser et al. 2013; Putzke et al. 2010; Shi et al. 2014; Trier 2008; Trier and Richter 2014; Whelan 2007; Whelan et al. 2013) and the other trying to infer value from the communicative content shared within these networks (i.e., sentiment analysis) (e.g., Cheung et al. 2012; Clemons et al. 2006; Jensen et al. 2013; Li and Hitt 2010; Mudambi and Schuff 2010). Yet, at the same time, there is evidence to suggest that invaluable insights could be gleaned from research that considers the structural properties and communicative content of social networks in tandem (see Butler et al. 2014; Chau and Xu 2012; Füller et al. 2014; Gasson and Waters 2013; Gray et al. 2011; Moser et al. 2013; Trier and Richter 2014). Therefore, in distinguishing between social graph and social text as constituent elements of social data, our proposed theory gives equal prominence to the two aforementioned research streams by embracing the structural properties and communicative content of social media.

[15] URL: http://www.forbes.com/sites/chuckjones/2013/12/03/ubs-analyst-milunovich-upgrades-apple-to-buy-with-650-price-target.

Second, our theory of social data is the first to bring clarity to plausible dimensions that could be incorporated into empirical predictive models for social media (see Figure 1). By deriving constituent dimensions of social graph (i.e., actor, action, activity and artifact) and social text (i.e., topic, keywords, pronoun and sentiment), we enlarge the pool of options for applying predictive analytics to big social data. Third, we demonstrate the applicability of our proposed theory through the construction of empirical predictive models that are invariant to the kind of social media platform (i.e., Facebook and Twitter) from which data is extracted and the type of corporate entities (i.e., financial performance of H&M and iPhone) to be predicted, be it companies or products. In this sense, our proposed theory of social data can be deemed as a cornerstone for future studies of predictive big social data analytics to build upon.

Last but not least, beyond predictive analytics, we believe that our proposed theory of social data can also aid in the generation of holistic frameworks for computational social science in general and big social data analytics in particular. So far, computational methods, formal models and software tools for big social data analytics have been largely confined to graph theoretical approaches (Gross and Yellen 2005) in the likes of social network analysis (Borgatti et al. 2009), which in turn is informed by the social philosophical approach of relational sociology (Emirbayer 1997). As far as we know, there is no other unified modeling approaches to social data that assimilates conceptual, formal, software, analytical and empirical domains (Mukkamala et al. 2013). Recent work (e.g., Vatrapu et al. 2014a, 2014b) has sought to outline an alternative approach to the predominant triad of relational sociology, graph theory and social network analysis, which are founded on associational sociology (Latour 2005), set theory and fuzzy set theory (Ragin 2000) as well as social set analysis (Mukkamala et al. 2014).

II.4.2. Implications for Practice

This paper should be of interest to practitioners for three reasons. First, our empirical results bear direct and indirect implications for companies. Naturally, a direct and obvious implication from this study is the proof that business performance can be predicted from big social data. By extracting and analyzing data from multiple social media platforms (i.e., Facebook and Twitter) to predict the financial performance of both companies (i.e., H&M) and products (i.e., iPhone), we are able to show that the predictive power of big social data is neither constrained by the social media platform nor the type of parameter to be predicted. For this reason, the indirect implications are that companies should proactively engage and strategically manage social media platforms in order to benefit from the strong correlations between social media interactions and sales performance. Second, by delineating social data into elements of social graph and social text, we provide companies with a schema of the elements to pay attention to on social media platforms. In order for companies to generate competitive advantage from social media, they must not only recognize the structural relationships within social networks, they must also value the opinions and sentiments embodied within social media content. Finally, this study is the first of its kind to take into account the existence of a time-lag from the moment a potential customer becomes aware of a product to the instance he/she decides to acquire it via a sales transaction when building empirical predictive models. In a way, this study highlights the importance of social media as an inexpensive forum for companies to continuously maintain product awareness in the minds of consumers.

II.4.3. Limitations

There are several limitations to the work reported here. First, we lack multiple cases to extensively evaluate and validate the overall prediction model. A second limitation is the emerging challenge for predictive analytics from social data associated with increasing sales in emerging markets such as China with its own unique social media ecosystem. By and large, the social media ecosystem of China

does not overlap with that of Western countries to which Facebook and Twitter belong. We suspect that the effect of non-overlapping social media ecosystems might be somewhat ameliorated for Veblen goods such as iPhones given the conspicuous consumption aspirations of a global middle class. This however remains an analytical challenge and restricts the predictive power of our H&M prediction model. A third limitation of the paper is that the theory of social data is limited to a cross-sectional framework of social data in terms of social graph (i.e., actors, actions, activities and artefacts) and social text (i.e., topics, keywords, pronouns and sentiments). As such, our theory of social data does not outline a process model, which might be more pertinent to predictive analytics. A fourth limitation arises from the representativeness of social media data. That said, as far as predictive analytics of real-world activities is concerned, social media datasets might be adequately representative as long as the basic premise of a social media action being a proxy for a user's attention to that particular real-world activity holds true. Our theory of social data will only cease to be valid if and when a user's social media action (such as a tweet about an "iPhone") is not a proxy for that user's attention towards the "iPhone" object. In our view, this fundamental disjunction between social media actions and real-world attention is the Achilles's Heel of predictive analytics with social data and might partially explain the spectacular drop in accuracy for once popular prediction models like the Google Flu Prediction System. A fifth and final limitation of our study, as far as our knowledge goes, is the lack of theoretical explanation for the empirical values for the time lags both in the nominal sense and the relative sense of divergence between Facebook and Twitter.

II.4.3. Future Work

For future work, we envision several projects that could spawn from this research as outlined below.

Going beyond the traditional and pre-dominant sentiment classification of social

text and towards domain-specific classifiers such as AIDA and HoE for predicting sales. This will require not only sophisticated computational linguistics methods and tools but also critical contributions from domain experts (e.g., for training datasets in the case of supervised machine learning algorithms).

Investigating other predictor variables such as socio-economic factors, confidence, trust, loyalty etc. Essentially. Moving towards "thick models" of human users and narrowing the social media user and real-world consumer gap for non-digital products and services.

Combining social media data with other online sources such as Google Trends or in-house data of enterprise systems such as ERP and CRM.

II. References

Asur, S. and Huberman, B. A. "Predicting the Future with Social Media," in Proceedings of the *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT) (Vol. 1), 2010, pp. 492-499.

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G. "Network Analysis in the Social Sciences," *Science* (323:5916), 2009, pp. 892-895.

Butler, B. S., Bateman, P. J., Gray, P. H. and Diamant, E. I. "An Attraction–Selection–Attrition Theory of Online Community Size and Resilience," *MIS Quarterly* (38:3), 2014, pp. 699-728.

Chau, M. and Xu, J. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), 2012, pp. 1189-1216.

Cheung, M. Y., Sia, C. L. and Kuan, K. K. "Is This Review Believable? A Study

of Factors Affecting the Credibility of Online Consumer Reviews from an ELM Perspective," *Journal of the Association for Information Systems* (13:8), 2012, pp. 618-635.

Clemons, E. K., Gao, G. G. and Hitt, L. M. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), 2006, pp. 149-171.

Emirbayer, M. "Manifesto for a Relational Sociology," *The American Journal of Sociology* (103:2), 1997, pp. 281-317.

Füller, J., Hutter, K., Hautz, J. and Matzler, K. "User Roles and Contributions in Innovation-Contest Communities," *Journal of Management Information Systems* (31:1), 2014, pp. 273-308.

Gasson, S. and Waters, J. "Using A Grounded Theory Approach to Study Online Collaboration Behaviors," *European Journal of Information Systems* (22:1), 2013, pp. 95-118.

Gray, P. H., Parise, S. and Iyer, B. "Innovation Impacts of Using Social Bookmarking Systems," *MIS Quarterly* (35:3), 2011, pp. 629-643.

Gross, J. L. and Yellen, J. *Graph Theory and Its Applications*, CRC press, 2005.

Hussain, A. and Vatrapu, R. "Social Data Analytics Tool," DESRIST 2014, *Lecture Notes in Computer Science* (LNCS), 8463(Springer), 2014, pp. 368–372.

Jensen, M. L., Averbeck, J. M., Zhang, Z. and Wright, K. B. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective," *Journal of Management Information Systems* (30:1), 2013, pp. 293-324.

Johnson, S. L., Faraj, S. and Kudaravalli, S. "Emergence of Power Laws in Online

Communities: The Role of Social Mechanisms and Preferential Attachment," *MIS Quarterly* (38:3), 2014, pp. 795-808.

Lassen, N., Madsen, R. and Vatrapu, R. "Predicting iPhone Sales from iPhone Tweets," in Proceedings of the *18th IEEE Enterprise Computing Conference* (EDOC 2014), Ulm, Germany, 2014.

Latour, Bruno (2005). *Reassembling the social an introduction to actor-network-theory*. Oxford New York: Oxford University Press. ISBN 9780199256044.

Lavidge, R. J. and Steiner, G. A. "A Model for Predictive Measurements of Advertising Effectiveness," *Journal of Marketing* (25:6), 1961, pp. 59-62.

Li, H. and Leckenby, J. "Examining the Effectiveness of Internet Advertising Formats," in D. Schumann & E. Thorson (eds.), *Internet Advertising: Theory and Research*, Lawrence Erlbaum Associates, 2007, pp. 203-224.

Li, X. and Hitt, L. M. "Price Effects In Online Product Reviews: An Analytical Model And Empirical Analysis," *MIS Quarterly* (34:4), 2010, pp. 809-831.

Moser, C., Ganley, D. and Groenewegen, P. "Communicative Genres as Organizing Structures in Online Communities–of Team Players and Storytellers," *Information Systems Journal* (23:6), 2013, pp. 551-567.

Mudambi, S. M. and Schuff, D. "What Makes A Helpful Online Review? A Study Of Customer Reviews on Amazon.Com," *MIS Quarterly* (34:1), 2010, pp. 185-200.

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Formal Model of Social Data," *IT University Technical Report Series*, TR-2013-169, 2013. [Available online at: https://pure.itu.dk/ws/files/54477234/ITU_TR_54472013_54477169.pdf, accessed October 14, 2014]

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Set Theoretical Approach to Big Data Analytics," in *Proceedings of IEEE Big Data 2014*, Anchorage, United States of America, 2014.

Putzke, J., Fischbach, K., Schoder, D. and Gloor, P. A. "The Evolution Of Interaction Networks In Massively Multiplayer Online Games. *Journal of the Association for Information Systems* (11:2), 2010, pp. 69-94.

Ragin, C. C. *Fuzzy-Set Social Science*, University of Chicago Press, 2000.

Sein, M., Henfridsson, O., Purao, S., Rossi, M. and Lindgren, R. "Action Design Research," *MIS Quarterly* (35:1), 2011, pp. 37-56.

Shi, Z., Rui, H. and Whinston, A. B. "Content Sharing in A Social Broadcasting Environment: Evidence From Twitter," *MIS Quarterly* (38:1), 2014, pp. 123-142.

Shmueli, G. and Koppius, O. R. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), 2011, pp. 553-572.

Trier, M. "Research Note-Towards Dynamic Visualization For Understanding Evolution Of Digital Communication Networks," *Information Systems Research* (19:3), 2008, pp. 335-350.

Trier, M. and Richter, A. "The Deep Structure of Organizational Online Networking–An Actor-Oriented Case Study," *Information Systems Journal* (Advance copy) 2014.

Vatrapu, R. "Understanding Social Business," In K. B. Akhilesh (ed.), *Emerging Dimensions of Technology Management*, New Delhi: Springer, 2013, pp. 147-158.

Vatrapu, R. and Suthers, D. "Intra-and Inter-Cultural Usability in Computer-Supported Collaboration," *Journal of Usability Studies* (5:4), 2010, pp. 172-

197.

Vatrapu, R. Cultural Considerations in Computer Supported Collaborative Learning," *Research and Practice in Technology Enhanced Learning* (3:2), 2008, pp. 159-201.

Vatrapu, R. K. "Explaining Culture: An Outline of a Theory of Socio-Technical Interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, Copenhagen, Denmark, 2010, pp. 111-120.

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings," in *Computational Social Science: Contagion, Collective Behaviour, and Networks*, Oxford: University of Oxford, 2014a, pp. 22-24. [available online at: http://cssworkshop.oii.ox.ac.uk]

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "Towards a Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Empirical Findings," in *Proceedings of the 5th Annual Social Media & Society International Conference 2014*, Toronto, Canada, 2014b.

Whelan, E. "Exploring Knowledge Exchange In Electronic Networks of Practice," *Journal of Information Technology* (22:1), 2007, pp. 5-12.

Whelan, E., Golden, W. and Donnellan, B. "Digitizing The R&D Social Network: Revisiting The Technological Gatekeeper," *Information Systems Journal* (23:3), 2013, pp. 197-218.

Paper III: Social Media Data as Predictors of Mikkeller Sales?

In the Proceedings of the 39th Symposium i Anvendt Statistik, University of Southern Denmark, Odense, Denmark, 23–24 January 2017.

Social media data as predictors of Mikkeller sales?

By

Niels Buus Lassen, Dep. of IT Management, CBS

Lisbeth la Cour, Dep. of Economics, CBS

Anders Milhøj, Dep. of Economics, KU

Ravi Vatrapu, Dep. of IT Management, CBS

III.1. Introduction

In recent years, social media data such as Twitter, Facebook and Google Trends data have proven promising as predictors for measures of economic outcomes of private firms. The main advantage of using social media data as predictors lies in the speed with which such data can be extracted and employed in the forecasting process. Once a firm has learned how to collect and pre-process their social media data, the information is available almost in real time and this implies that such data in combination with a good predictive model will provide a very useful tool for the management of the firm.

When working with social media data the concept of 'Big data' often comes to people's minds. In our case this is only partly true: we do work with large amounts of social media data, but once they have been pre-processed, we end up as many studies in the literature using quite simple dynamic regression models based on rather few time series observations. Hence the whole distinction between 'tall', 'fat' and 'huge' data as suggested in Doornik & Hendry (2014) becomes of less relevance. Ideally, if we were able to get economic performance data for a firm at

a high frequency like the daily frequency, we would move closer to a situation where a more automatic model selection procedure would be relevant.

The novelty of the present paper is a predictive model for the total sales of Mikkeller using data at a monthly level. With these data we are allowed to be more precise when it comes to specification of the lag-structure in the dynamic regression model. Also we look into the importance of the data-preparatory work – in our case an unobserved component filtering of the data prior to regression modeling - on the social data proves to be for the final model and finally, we investigate the predictive power of types of social media data that have not been used as predictors before for a brewing company: Google shopping and YouTube data.

III.2. Briefly on the existing literature

The idea of using social media data as predictors for e.g. company sales is not new. When it comes to model building, various experiments have been conducted and a summary of around 40 articles covering the time period 2005 – 2015 can be found in Buus Lassen et al (2017). For the present purpose the most interesting observations from these studies are that 1) almost 50% of the studies use some kind of regression model as their predictive model, 2) the range of social data types studied seem to cover Facebook, Twitter, Google Trends , Instagram, Tumblr, blogs and Youtube.

Theoretically, the argument for considering social data activity as predictors for sales obtains support from e.g. the AIDA model mentioned in Buus Lassen et al (2014). AIDA means *Awareness, Interest, Desire and Action* and refers to stages in a sales process. If social media data help increase the attention or can be considered a proxy for attention towards a product then it may also affect the final decision about buying. It is the general perception that more attention will increase

sales even if the attention is negative.

When it comes to the specification of a set of predictive models we follow the literature and limit ourselves to the class of dynamic regression models. In these models we will have sales as our dependent variable and the different social data as suggested regressors. The reason why it is of interest to study social data regressors from different social media and search sources lies in the different ways such media are used. Google searches have proven to be the best social data for predicting sales. We call the Google data unpolished with a good connection to people's brains. Facebook data are polished, because people tend to display success and not failures on this social data. Twitter data are better than Facebook data for sales modeling, because Twitter data are less polished. But Google data are still beating Twitter data for sales modeling, because Google data are unpolished.

When building predictive models, the data frequency is of high importance. The higher the data frequency the more room is given for the researcher to build dynamic regression models with the aim of eventually forecasting economic performance measures such as sales. For many private companies sales data will only be available from the accounting data at a quarterly frequency (official balance sheets). This will limit the number of observations, and characteristics such as seasonal patterns may be more difficult to extract. Monthly data are much better as more observations will usually be available but at this frequency regular patterns over the week will still be impossible to discover. If possible to get, data at a daily frequency would be very well suited from all perspectives but are rarely available. In the present study we are able to work with monthly sales data for Mikkeller, a Danish micro brewery which has activities all over the world.

III.3. The data and methodology

In order to build a predictive model for Mikkeller's sales we use data from

Mikkellers accounting system combined with Google Trends, Google shopping and YouTube data. The social data has been collected from the free-access numbers available on the respective WEB-pages. We have searched for the word 'Mikkeller'. The free data from Google that we use are indexed such that they will vary between 0 and 100[16]. The time span of the study has been limited by our access to historical sales data and also the frequency of the data is reflecting our access to Mikkeller data. In the end we have a sample of monthly data that covers Januar 2014 to September 2016. Prior to analysis we index the sales data such that the max value becomes equal to 100. This transformation does not affect significance results later in the modeling process.

III.3.1 Pre-processing methodology

Our first considerations when it comes to data preparatory work concerns whether to use simple transformations of the series or just the raw series themselves. From a graphical inspection of total sales and the log of total sales it seems that using the log of sales may offer a slight statistical advantage as the variance seems more constant over the sample period than for the raw series, see Figures 1A and 1B.

With respect to the sales data we are checking the stationarity properties of the time series by means of an ACF graph. Stationarity is preferable for a regression model although stationarity may be of minor importance when the purpose of the model is forecasting.

The social data may consist of different components that we would expect to have different predictive value. Prior to including our social data time series as explanatory factors in our regression models we have the possibility to split them into a trend component, a seasonal component and an irregular component using

[16] It is possible to get the actual number of searches but they are not available for free.

classical times series techniques for unobserved components models (ucm). Our prior is that the irregular component will contain the most valuable information for predictive purposes as this component will capture special events, that creates a lot of attention towards the firm and its products. We also estimate models that use the social data in their 'raw' form without the ucm pre-processing for comparison reasons.

III.3.2 Unobserved Component Models

An unobserved component model, UCM, decomposes the observed series y_t into a sum of many components, as for instance

$y_t = \mu_t + \varepsilon_t$
$\mu_t = \mu_{t-1} + \eta_t$

Here the series μ_t is understood as the level of the series; but this level is unobserved. Only the series y_t which is affected by some noise or irregularities is observed. This noise series, ε_t, could in technical applications be measuring errors. But in this presentation the series ε_t is used as the irregular component which consists of special events happening to the series at time t which are not a part of the underlying level μ_t. In this paper these irregular components which are estimated for the observed sales series and for the three social data series are used in a usual regression/time series model in order to see if the social data series have any impact to the sales data in a setup where all usual time series variation for each series is accounted for by the unobserved components.

This basic formulation could be extended by trends and seasonality, and various forms for introducing autocorrelation in the model formulation also exist. A trend component has the form:

$\beta_t = \beta_{t-1} + \xi_t$

and the seasonal component is defined in a way so it does not affect the level component:

$S_t = -(S_{t-1} + ... + S_{t-11}) + \zeta_t$

In total these ideas lead to the model:

$y_t = \mu_t + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

where also an autoregressive term for the irregular series is included.

All remainder terms, ε_t, η_t, ξ_t and ζ_t, are assumed to be mutually independent white noise series. Their variances could be estimated; the larger this component variance the more volatile the component. But it is also possible to fix this variance to the value zero which gives a constant component, e.g. a model with fixed seasonal dummies is found if $var(\zeta_t) = 0$. But if $var(\xi_t) > 0$ the trend is allowed to vary over time which is a very flexible feature!

The parameters of these models, the variances and the AR(1) parameter, and the component values could be estimated by the Kalman filter. This gives an algorithm for successive calculation of the unobserved components at time t conditioned on previous observations y_{t-i} i = 0, .. , t-1. The Kalman filter is useful if prediction is the purpose of the analysis as the algorithm does not include future observations y_{t+i}. A further smoothing estimation, where all available information is used when estimating the unobserved components at any time t, also exist. In this paper this method will be used.

Our hypothesis when it comes to the UCM components is that they probably will have most potential if the data frequency is high. With our monthly data the idea may still be applicable but there is a danger that the temporal aggregation level will make it more difficult to find the type of effects we are looking for.

III.3.3 The regression models

In order to specify a predictive model, we direct our focus to the class of dynamic regression models. Unfortunately, even with monthly data we are left with rather few observations which will limit our possibility to work with both complex lag structures and many non-linear terms like power expressions and interactions.

The primary model equations we use are of the type:

$$y_t = \beta_0 + \gamma y_{t-1} + \beta_1 x1_{t-1} + \beta_2 x2_{t-1} + \ldots + \beta_k xk_{t-1} + \varepsilon_t \quad t = 1, \ldots, T \quad (1)$$

where y is sales, xj is a social data measure and the sub scripts, t - 1, indicate that only lagged values of sales and social media data are used as predictors. The basic model can be extended by allowing for more than one lag of each xj variable but due to the limited number of observations for the estimation period the more predictors we include the fewer lags we can afford to consider. Also by including the lag of y in the equation we actually consider a longer lag structure of x but with a specific exponentially decaying pattern in the effects over time. Hence our preferred initial specification as provided by equation (1) will include the lag of y and only one lag for each additional explanatory factor. The error term, ε_t, is assumed to fulfill the standard assumptions for OLS estimation.

We also provide empirical evidence based on a model of the type:

$$y_t = \beta_0 + \gamma y_{t-1} + \beta_1 x1_{t-j} + \varepsilon_t \quad t = 1, \ldots, T \quad (2)$$

where the error term, ε_t, again is assumed to fulfill the standard assumptions for OLS estimation. Choosing to include just the j'th lag may be based on more empirical arguments. Also y_{t-1} may be left out of equation (2) if that seems to make more sense.

It is difficult to judge the predictive performance of a specific forecasting model unless we have some benchmark to compare to. For sales of individual companies

there is no general guideline in the literature on how to choose such a model, so we will argue for our choice in the following way: we want a benchmark model that is simple, that seem to capture some of the apparent time series properties in our data and that do not contain exogenous explanatory factors. In this study we will suggest two such model 1) a simple AR(1) model[17] as suggested by the standard identification procedure from classical time series analysis (see Figures 2A and 2B.

In addition to this model we also consider a model that includes the first lag of log sales but also a December dummy and a time trend. The December dummy is at first hand negative but as the sales numbers are at the time of production the low December values are due to low sales in January or later months.

Benchmark 1: $y_t = \beta_0 + \gamma y_{t-1} + \varepsilon_t \quad t = 1, \ldots, T$ (3)

Benchmark 2: $y_t = \beta_0 + \gamma y_{t-1} + \beta_1 D\_December_t + \beta_2 Trend_t + \varepsilon_t \quad t = 1, \ldots, T$ (4)

In the end, as our model is a forecasting model, we need to split the sample into a training and a test part in order to assess the out-of-sample forecasting properties, see e.g. Hyndman & Athanasopoulos (2014). With our short sample we retain only the last 3 month of the sample for the test part, i.e. July – September 2016. This leaves us with an estimation or training sample of 30 observations: January 2014 – June 2016. We will provide an out-of-sample forecast results based on a series of 1 step ahead predictions for July – September 2016. Evaluations will be based on graphs comparing actual sales to predicted sales for the selected models and also by numerical measures like RMSE and MAE.

III.4. Descriptive statistics

In this section we will provide a series of graphs and tests that will help us illustrate

[17] Because we have no indication of non-stationarity of our sales series, we go for the AR(1) specification instead of a random walk which is often used as a benchmark in the exchange rate literature.

the basic time series properties of the data and support us in arguing for the transformations we decide to use in our models. We start by showing the development over time in both the sales and the log of sales, see figures 1A and 1B. The first impression of the development of the series is that there seems to be some indication of an upwards trending behavior. An alternative to this interpretation could be an interpretation of a non-trending series but with a level shift upwards. We will start by considering the first case of an increasing trend as it makes it unnecessary for us to decide on an exact timing of a level shift. But it may be worthwhile to keep this second option in mind for future studies. When comparing the graphs of sales and log of sales it is not clear which one to prefer. In the end we decided to go for the log of sales as this is often done for longer time series where the variation increases with increases in the level. Also focusing on the logs will allow for an interpretation in percentage terms when it comes to the analysis of the regression models.

Figure 1A: Total Sales

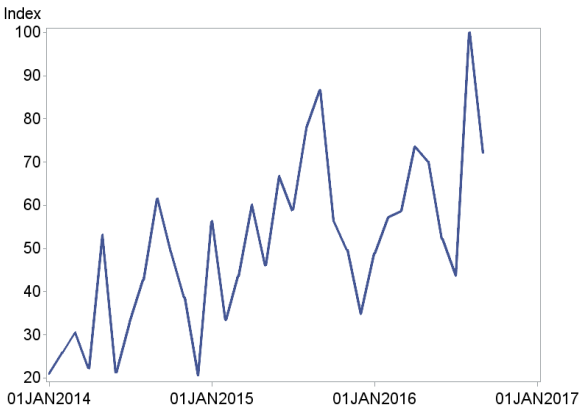
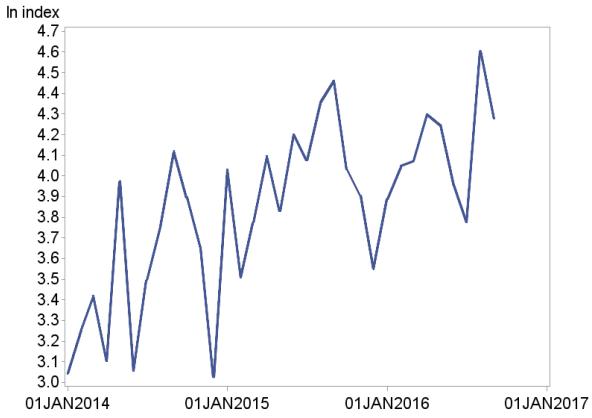


Figure 1B: Log of Total Sales



When taking a closer look at the time series properties of the series it seems that a decision of treating this series as stationary would be a good starting point. The ACF graph clearly supports this conclusion as the 1st order autocorrelation coefficient is 0.41[18].

[18] Also an ADF test of non-stationarity of the series supports a conclusion of stationarity. With a trend in the equation of this test we reject at the 10% level the null of a unit root with p-values of 0.001 and 0.059 for zero and 1 lagged differences in the equation, respectively. However with our very short sample period this test may not be too reliable. At least it does not contradict our impression from the ACF graph.

Figure 2A: ACF for log of Total Sales

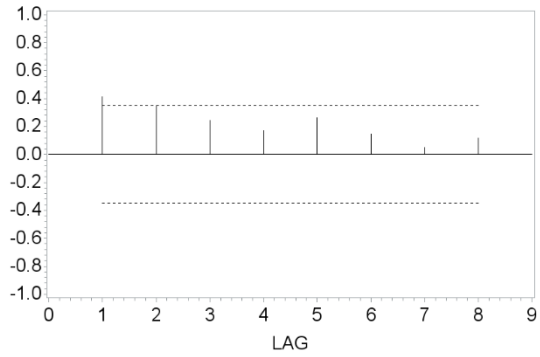
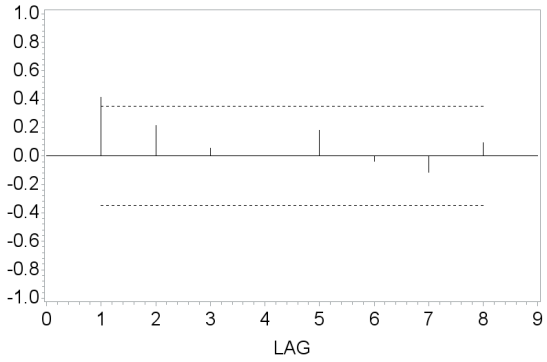


Figure 2B: PACF for log of Total Sales



To get a first impression of potential linkages from social data activity to the Mikkeller sales we plot each social data activity in a time series plot together with the log of sales. As seen in figures 3A – 3C none of the social data series seem to follow the sales very closely – not even if some lagging is considered. To gain further knowledge about the correlation behavior we also provide a matrix plot of log sales and the potential regressors. There is for some of the variables a vague pattern that would support a correlation different from zero but the general picture is not too promising.

Figure 3A: Log Sales and Google searches

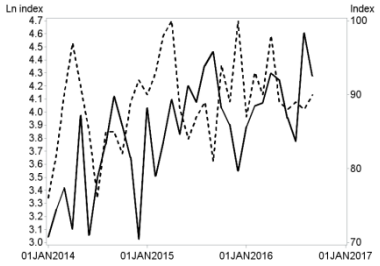


Figure 3B: Log Sales and Google shopping

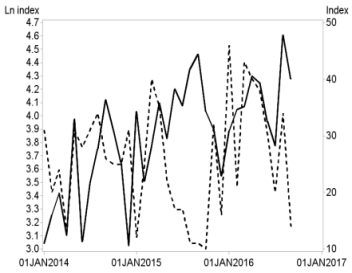
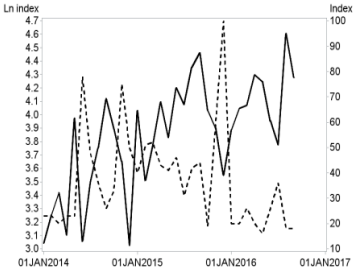
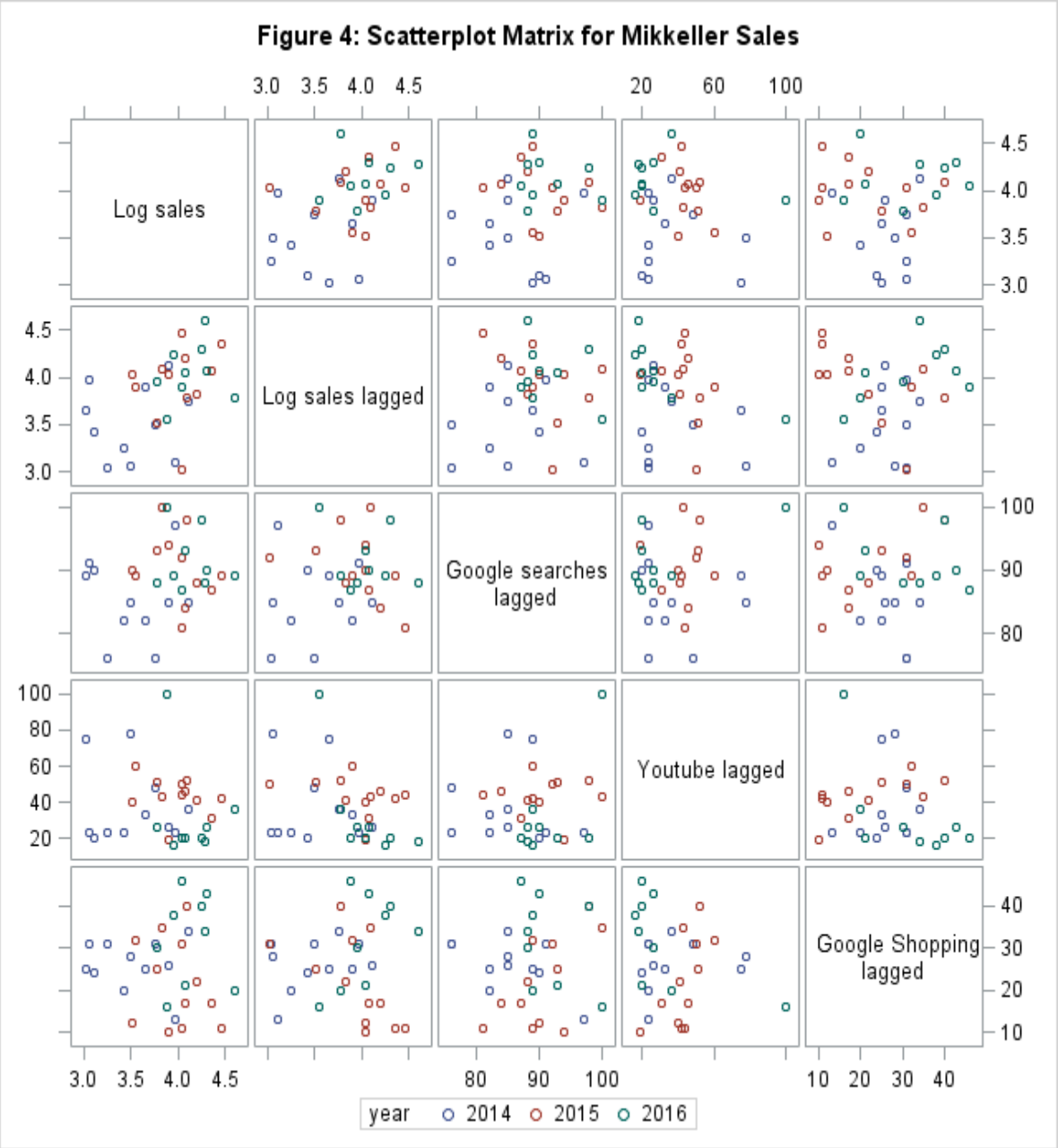


Figure 3C: Log Sales and Youtube



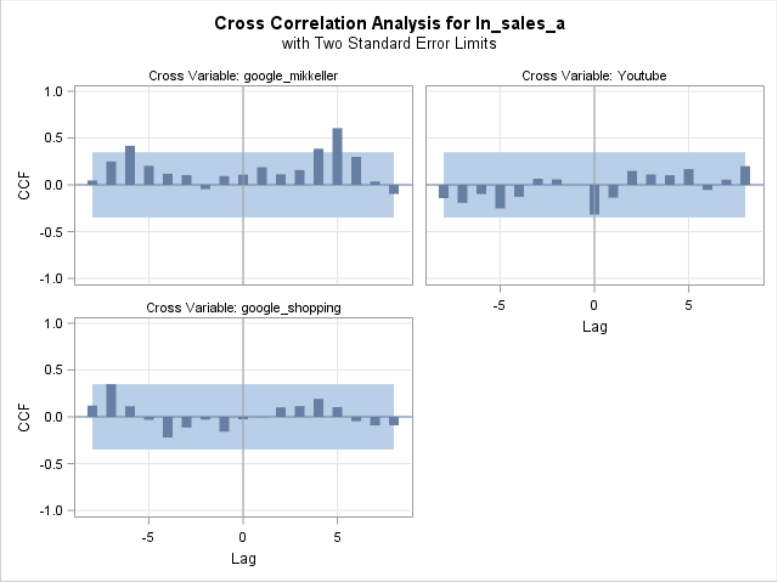
Note: in Figures 3A – 3C the solid curve represents the sales.



Next in order to get some more knowledge about possible lagged effects, in Figure 5 we present the cross correlations between the indexed sales variable and each of the social activity variables. The cross correlations are constructed in such a way that a positive number, s, on the horizontal axis implies a correlation between sales at time t and the social variable s periods prior to the present one. It seems from these graphs that we cannot expect much explanatory power from including these social media variables as regressors in a predictive model for sales. The only

significant spike is for Google searches with a lag of five. If many of the customers buy their quantities at irregularly rather than on a smooth continuous basis it may make sense to include the lag 5 variable in a regression model. It is, however, also possible that this lag becomes significant for more random reasons due to our short sample period. An out-of-sample test of the model may help here.

Figure 5



Finally, we display the summary statistics of sales and the four series of social media attention:

Table 1: Descriptive summary statistics.

Variable	N	Mean	Std Dev	Minimum	Maximum
Log sales	33	3.8423177	0.4212766	3.0264408	4.6051702
Google searches	33	88.9393939	5.9315860	76.0000000	100.0000000
Youtube	33	37.1818182	19.6793905	16.0000000	100.0000000
Google Shopping	33	25.8484848	9.9784806	10.0000000	46.0000000

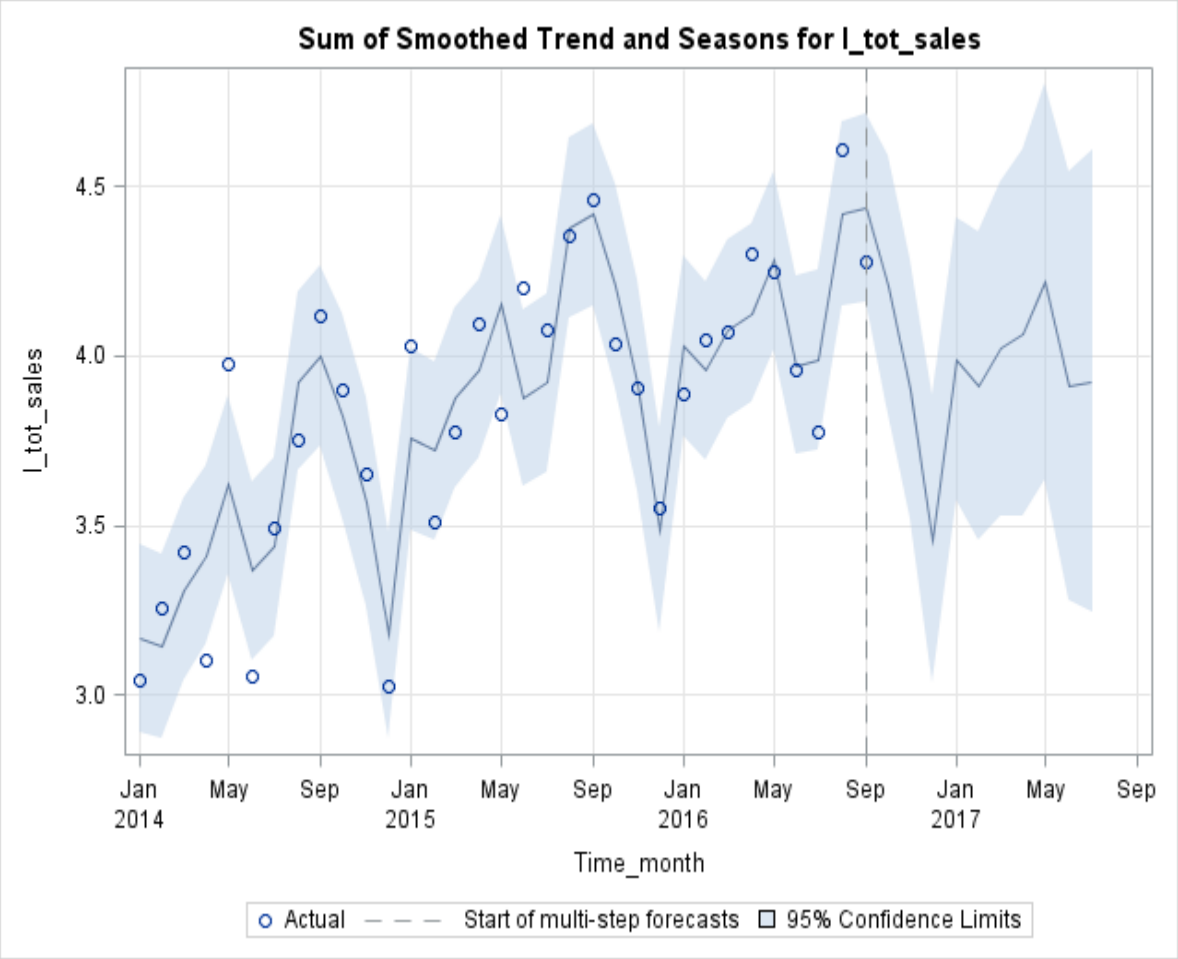
III.5. Unobserved components models for the sales series and the social media series.

For the log transformed, indexed sales series the resulting model is:

$y_t = \mu_{t-1} + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

The variance in the seasonal dummy series is fixed to zero, meaning that the dummy variables are constant. The trend variance varies and this allows the trend to be significantly positive for the first year or two and then the trend is zero. This changing trend is clearly seen at the fit plot even if the seasonal component is also present at the plot.

Figure 6



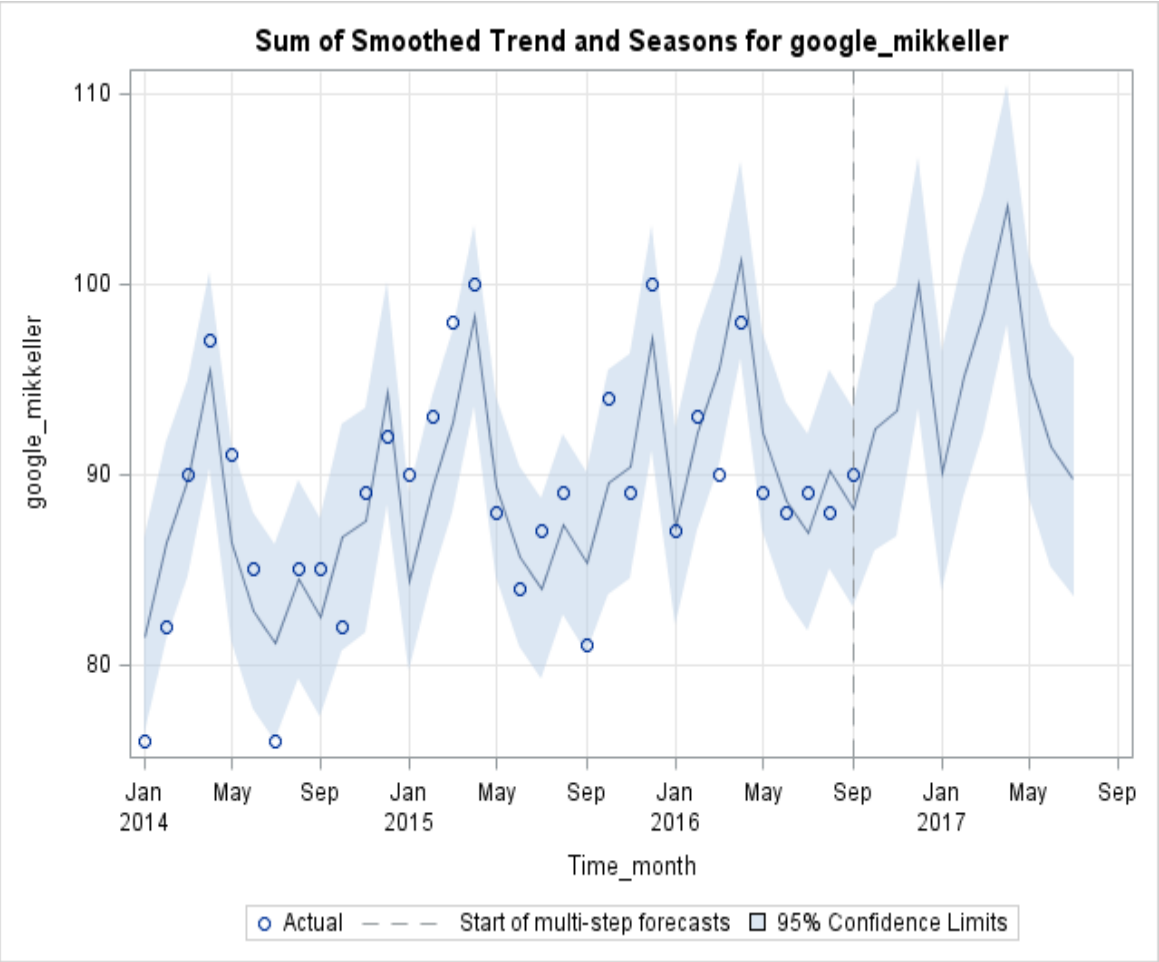
For the Google series the resulting model is:

$y_t = \mu_t + \beta_t + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

Again the variance in the seasonal dummy series and now also in the trend series are fixed to zero, meaning that the trend and the seasonal dummy variables are constant. The series has a constant upward trend at 0.25 each month, which is

significantly positive. This means that the interest in google searching for the word "Mikkeller" is steadily increasing. The trend is clearly seen at the model plot in figure 7.

Figure 7



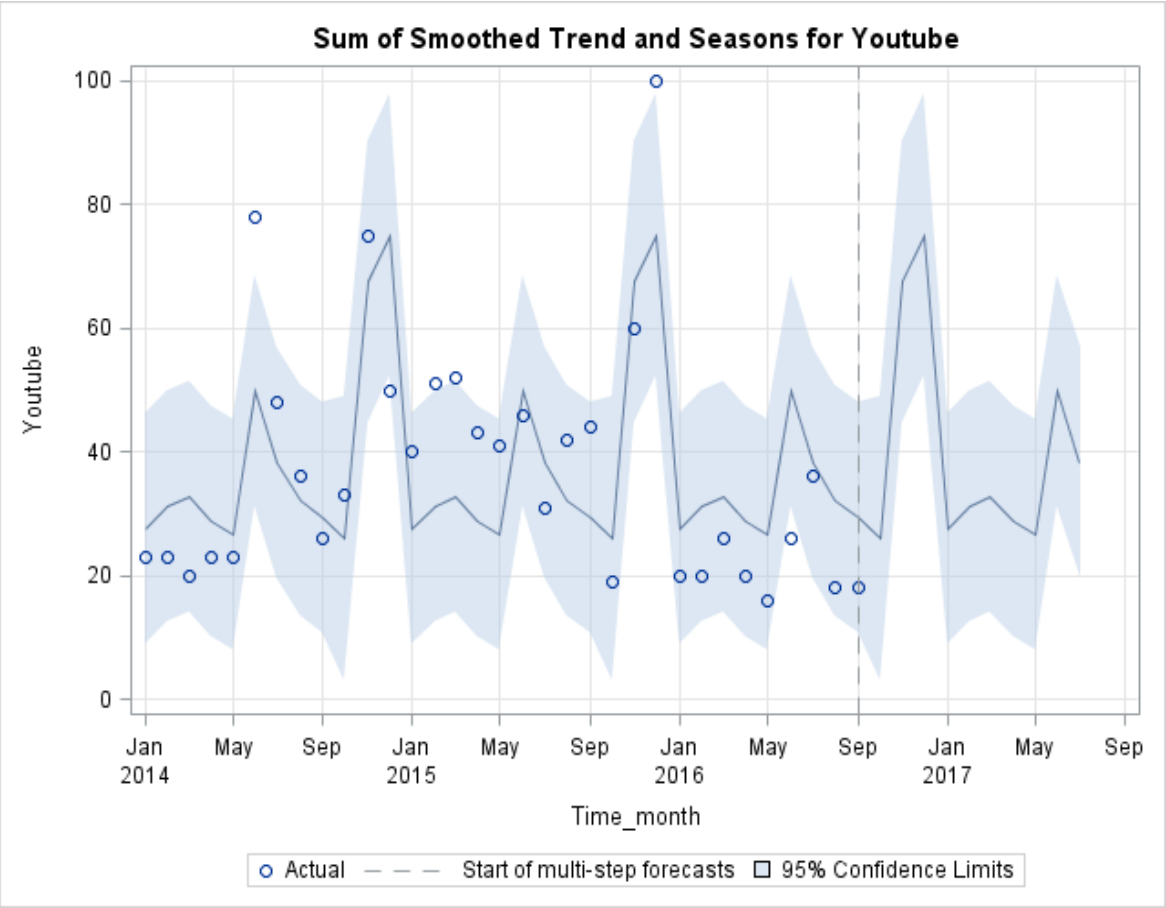
The seasonal component tells that Google searching for the word "Mikkeller" has peaks every year in April and December.

For the Youtube series the AR(1) term is insignificant and no trend is present. The resulting model is

$y_t = \mu_t + S_t + \varepsilon_t$

where all variances are fixed at zero. The model plot tells that the series has a clear seasonal component with large values in the months of November and December.

Figure 8



For the series of Google Shopping the resulting model is even more simple as the seasonal component is insignificant.

$y_t = \mu_t + \varepsilon_t$

III.6. Results of predictive modeling at the monthly frequency.

We now consider various specifications for models that contain social media activity data and/or their lags as explanatory factors as suggested by the main equation (1) but also by the additional equation (2).

As our main purpose it to determine a model that can produce 1 step ahead forecasts out-of- sample, we will as a starting point not allow for contemporaneous regressors in the models. This may be a limitation when we are working with monthly data as social media activity in the beginning of a month may affect sales

already in the same month. Another possibility is that it actually takes several months for us to see a reaction if each customer only send an order with months in between.

III.6.1.1 In sample regression results.

Table 2: Regression results for Log Sales before ucm.

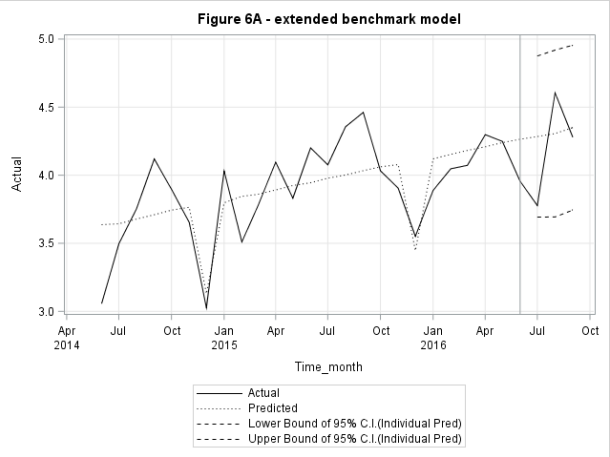
Variables	Benchmark 1 AR(1)	Benchmark 2 extended AR(1)	Model (1) Extended	Model with more lags	Model with lag 5
Intercept	2.57*** (0.77)	3.40*** (0.61)	4.19*** (1.38)	3.69** (1.47)	1.18** (0.55)
Lagged sales	0.34* (0.20)	0.02 (0.17)	0.03 (0.24)	-	-0.17 (0.12)
December	-	-0.65*** (0.19)	-0.70*** (0.23)	-	-0.42*** (0.13)
Trend	-	0.03*** (0.01)	0.03*** (0.01)	-	0.02*** (0.01)
Google searches, lag1	-	-	-0.01 (0.01)	0.01 (0.01	-
Youtube, lag1	-	-	0.00 (0.00)	-0.01* (0.00)	-
Google shopping, lag1	-	-	0.00 (0.01)	-0.00 (0.01)	-
Google searches, lag2	-	-	-	-0.01 (0.01)	-
Youtube, lag2	-	-	-	0.00 (0.00)	-
Google shopping, lag2	-	-	-	0.00 (0.01)	-
Google searches lag5	-	-	-	-	0.03*** (0.01)
Adj. R square	0.08	0.48	0.43	-0.06	0.79
# observations	25	25	25	25	25

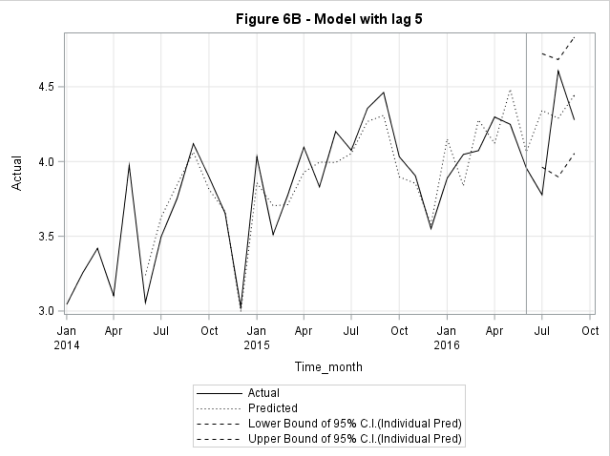
Note: the estimation sample has been restricted such that it is the same for all specifications even though models with fewer lags could have used more observations. Note2: Standard errors in parentheses. Significance at 10%: *, 5%:**, 1%: ***.

The basic massage from the table is that much of the variation in the log of sales can be explained by deterministic terms like a seasonal December dummy and a trend. The table also reveals that it is difficult based on the present sample to find significant effects from the social media activity variables. The most promising suggestion is to include the lag 5 of the Google searches in the model. This factor becomes very significant and also the model overall reaches an R square close to 80% in that case. The economic interpretation of including the lag 5 term is, however, more uncertain but may relate to lagged buying behavior from some customers.

III.6.1.2 Out-of-sample predictive power?

We predict the log of sales for the time period July 2016 to September 2016. First we show a set of graphs that compares such prediction for the actual values. We show graphs for the extended benchmark model and for the model including lag 5 in the last column of table 2.





From these graphs it is evident that not many of the movements in sales are captured by the benchmark model. Also the confidence bands for the prediction are quite wide and the actual values are inside the band. The 'lag 5'-model are capturing the movements in sales much better in sample but out of sample the big swings in July and August 2016 are not captured very well. The actual value for July is in fact outside the confidence bounds.

In table 3 we show some numerical measures for the forecasting performance of the models from table 2. We have chosen just to focus on a few measures and some of the more commonly used ones: MAE (mean absolute error) and RMSE (root mean squared error)[19].

Table 3: Summary measures on predictive power.

Summary measure	Benchmark 1 AR(1)	Benchmark 2 extended AR(1)	Model (1) Extended	Model with more lags	Model with lag 5
MAE	0.344	0.292	0.315	0.381	0.350
RMSE	0.449	0.342	0.363	0.444	0.385

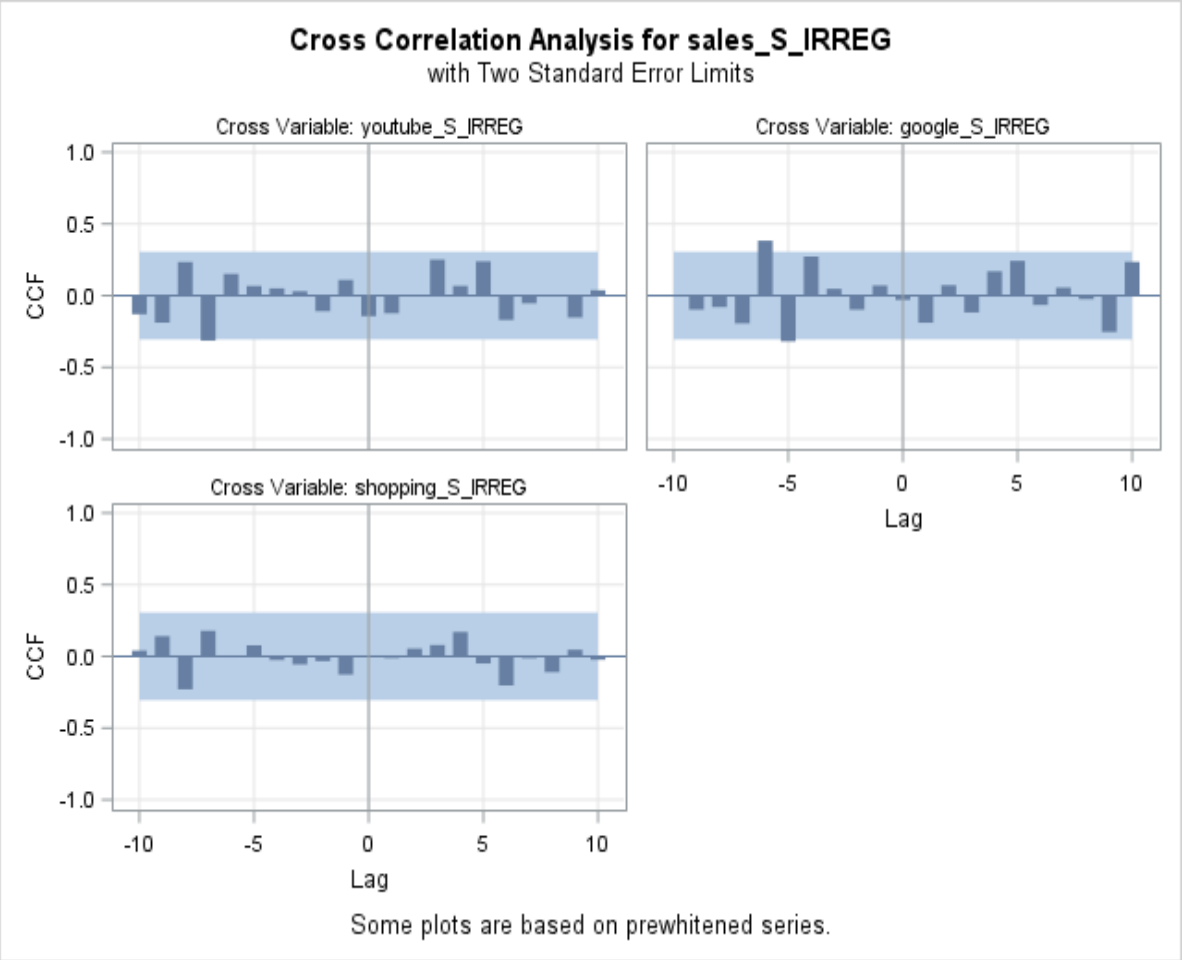
Note: In all cases the numbers have been calculated based on the 3 months of July, August and September 2016.

The numbers in table 3 also indicate that the extended benchmark model performs

[19] For formulas on how to calculate these measures , please consult e.g. Hyndman & Athanasopoulos (2014)

better when it comes to forecasting out-of-sample in our case. The model that performed best in-sample lies in the middle.

The results of this forecasting exercise may be due to the few months on which we base it and may also be a result of the very volatile behavior in this period. Somehow it seems that maybe some customers have decided to postpone their orders from July to August.



III.6.2. Predicting the irregular component for the sales series by the irregular components for the social media series. Initial investigations.

As an alternative to the attempts elsewhere in this paper, in this section we try to model the relations among the irregular components for all four series. This seems to be a fruitful way to go as the irregular components are cleaned from all trends, level shifts and seasonal variation. The relation between the sales on the left hand

side and the three social media series on the right hand side is however insignificant even if lags or leads are considered.

This is easily seen by the cross correlation functions; see the plots for the series S_IRREG above. Two of the irregular components include an autoregressive term by construction which means that these two series are prewhitened before calculating the cross correlations.

It is clear that no significant relations are found at these plot and we will not pursue this modeling further at the present stage.

III.7. Summary and conclusion

In this paper we have pursued our idea of applying a preparatory ucm model to both regressors and regressand to determine a forecasting model for the monthly sales of the Danish micro brewery Mikkeller. Our modeling attempts were mainly unsuccessful as the ucm modeling did not lead to any significant regression model based on regressors being activity from Google trends, Youtube and Google Shopping. Also when following a more traditional strategy without the preparatory ucm modeling, the benchmark model that contained a trend and a December dummy seemed to perform the best even though we found some support for a lag-5 effect from Google Trends. Much of our lack of success with the present modeling may be due to the fairly short sample period that consisted of monthly data from January 2014 until September 2016. Therefore future studies building on the same idea but with access to longer and maybe even more high frequent sample periods may prove more successful.

III. References

Buus Lassen, N., la Cour, L., Vatrapu, R. (2017), 'Predictive Analytics with Social Media data' in Sloan & Quan-Haase ed. *The SAGE Handbook of Social Media Research Methods*, Chapter 20, pp 328-341

Buus Lassen, N., Madsen, R. and Vatrapu, R. (2014). 'Predicting iPhone Sales from iPhone Tweets', Conference Paper, *2014 IEEE International Enterprise Distributed Object Computing Conference.*

Buus Lassen, N., Vatrapu, R., la Cour, L., Madsen, R. and Hussain, A.(2016), 'Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 241-256

Doornik & Hendry (2014). 'Statistical Model Selection with 'Big Data', *Department of Economics Discussion Paper Series*, University of Oxford, #735.

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts: https://www.otexts.org/fpp/

Paper IV: Predictive Analytics with Social Media Data

In SAGE Handbook of Social Media Research Methods, chapter 20, pages 328–341, 1st edition, SAGE Publications.

Predictive Analytics with Social Media Data

Niels Buus Lassen, Lisbeth la Cour, and Ravi Vatrapu

This chapter provides an overview of the extant literature on predictive analytics with social media data. First, we discuss the difference between predictive vs. explanatory models and the scientific purposes for and advantages of predictive models. Second, we present and discuss the foundational statistical issues in predictive modelling in general with an emphasis on social media data. Third, we present a selection of papers on predictive analytics with social media data and categorize them based on the application domain, social media platform (Facebook, Twitter, etc.), independent and dependent variables involved, and the statistical methods and techniques employed. Fourth and last, we offer some reflections on predictive analytics with social media data.

IV.1. Introduction

Social media has evolved into a vital constituent of many human activities. We increasingly share several aspects of our private, interpersonal, social, and professional lives on Facebook, Twitter, Instagram, Tumblr, and many other social media platforms. The resulting social data is persistent, archived, and can be retrieved and analyzed by employing a variety of research methods as documented in this handbook (Quan-Haase & Sloan, Chapter 1, this volume). Social data analytics is not only informing, but also transforming existing practices in politics, marketing, investing, product development, entertainment, and news media. This chapter focuses on predictive analytics with social media data. In other words, how

social media data has been used to predict processes and outcomes in the real world.

Recent research in the field of Computational Social Science (Cioffi-Revilla, 2013; Conte et al., 2012; Lazer et al., 2009) has shown how data resulting from the widespread adoption and use of social media channels such as Facebook and Twitter can be used to predict outcomes such as Hollywood movie revenues (Asur & Huberman, 2010), Apple iPhone sales (Lassen, Madsen, & Vatrapu, 2014), seasonal moods (Golder & Macy, 2011), and epidemic outbreaks (Chunara, Andrews, & Brownstein, 2012). Underlying assumptions for this research stream on predictive analytics with social media data (Evangelos et al., 2013) are that social media actions such as tweeting, liking, commenting and rating are proxies for user/ consumer's attention to a particular object/ product and that the shared digital artefact that is persistent can create social influence (Vatrapu et al., 2015).

IV.2. Predictive Models vs. Explanatory Models

At the outset, we find that the difference between predictive and explanatory models needs to be emphasized. Predictive analytics entail the application of data mining, machine learning and statistical modelling to arrive at *predictive models* of future observations as well as suitable methods for ascertaining the *predictive power* of these models in practice (Shmueli & Koppius, 2011). Consequently, predictive analytics differ from explanatory models in that the latter aims to: (1) draw statistical inferences from validating causal hypotheses about relationships among variables of interest, and; (2) assess the explanatory power of causal models underlying these relationships (Shmueli, 2010). This crucial distinction between explanatory and predictive models is best surmised by Shmueli & Koppius (2011) in the following statement: "whereas explanatory statistical models are based on underlying *causal relationships between theoretical constructs*, predictive models rely on *associations between measurable*

variables" (p. 556). For example, in political science, explanatory models have investigated the extent to which social media platforms such as Facebook can function as online public spheres (Robertson & Vatrapu, 2010; Vatrapu, Robertson, & Dissanayake, 2008) in terms of users' interactions and sentiments (Hussain, Vatrapu, Hardt, & Jaffari, 2014; Robertson, Vatrapu, & Medina, 2010a,b). On the other hand, predictive models in political science sought to predict election outcomes from social media data (Chung & Mustafaraj, 2011; Sang & Bos, 2012; Skoric, Poor, Achananuparp, Lim, & Jiang, 2012; Tsakalidis, Papadopoulos, Cristea, & Kompatsiaris, 2015).

Distinguishing between explanation and prediction as discrete modelling goals, Shmueli & Koppius (2011) argued that any model, which strives to embrace both explanation and prediction, will have to trade-off between explanatory and predictive power. More specifically, Shmueli & Koppius (2011) claim that predictive analytics can advance scientific research in six scenarios:

(1) generating new theory for fast-changing environments which yield rich datasets about difficult-to-hypothesize relationships and unmeasured-before concepts; (2) developing alternate measures for constructs; (3) comparing competing theories via tests of predictive accuracy; (4) augmenting contemporary explanatory models through capturing complex patterns which underlie relationships among key concepts; (5) establishing research relevance by evaluating the discrepancy between theory and practice; and (6) quantifying the predictability of measurable phenomena.

This chapter discusses predictive modelling of (big) social media data in social sciences. The focus will be entirely on what is often referred to as predictive models: models that use statistical and/or mathematical modelling to predict a phenomenon of interest. Furthermore, the focus will be on prediction in the sense

of forecasting a future outcome of the phenomenon of interest as such predictions are the ones that have so far received most attention in the literature. To illustrate the concepts, models, methods and evaluation of results we use examples from economics and finance. The general principles are, however, easily employed to other social science fields as well, for example, marketing. The concepts and principles that this section discusses are of a general nature and are informed by Hyndman & Athanasopoulos (2014) and Chatfield (2002). This chapter does not discuss applicable software solutions. However, it is worth mentioning that there exist quite a few software packages with more or less automatic search procedures when it comes to model specification. A few ones are, for example, SAS, SPSS and the Autometrics package of OxMetrics.

IV.3. Predictive Modelling of Social Media Data

When performing predictive analysis on social media data researchers often have to make a lot of decisions along the way. Examples of the most important decisions or choices will be discussed in the sections below.

IV.3.1. The phenomenon of interest and the type of forecasts

Quite often the focus will be on a single outcome (univariate modelling – one model equation) where the goal is to derive a prediction or forecast of, for example, sales in a company or the stock price of the company. In some cases, more than one outcome will be of interest and then a multivariate approach in which more than one relationship or model equation is specified, estimated, and used at the same time is worth considering. From now on let us assume that the phenomenon of interest is sales of a company and the social media data are among the factors that are considered as explanatory for the outcome. The discussion will then relate to the univariate case. At this stage, a decision is also necessary in relation to the data

frequency. Is the predictive model supposed to be applied to forecast monthly sale, quarterly sales or sales of an even higher frequency like weekly or daily?

IV.3.2.The data

Once the phenomenon of interest is identified, decisions concerning the data to be used have to be made. Data can be of different types: time series (e.g. sales per month or sales per day), cross sectional (e.g. individuals such as customers, for a given period in time) or longitudinal/panel (a combination of the former two such as a set of customers observed through several months). Predictive models can be relevant for all these types of data and many of the basic principles for analysis are quite similar. In the remaining parts of this section, for simplicity the focus will be on time series only.

As social media data have been growing in volume and importance during the last 10 years, in some cases the final number of observations for modelling may be rather limited as the dependent variable may reflect accounting and book-keeping and be relatively low-frequency like monthly or quarterly in nature. If this is the case, there may be a limit to how advanced models can be used. In other cases, daily data may be available and more complex models may be considered.

The frequency of the data is also important for model specification itself. With more high frequency data, a researcher may discover more informative dynamic patterns compared to a case with less frequent data. Consider a case where sales of a company need to be forecasted. If the reaction time from increased activity on the Facebook page of the company to changes in sales is short (e.g. just a couple of days) then if sales are available only on a monthly basis the lag pattern between explanatory factors and outcome may be difficult to identify and use.

In many cases there will be a large set of potential explanatory factors that may be included in various tentative model specifications. Social media data may be just a

part of such data and it will be important to also include other variables. The quality as well as the quantity of data is very important for building a successful predictive model.

IV.3.3. Social media data and pre-processing

When researchers consider using social media data for predictive purposes, at the outset the social media data will be collected at the level of the individual action (e.g. a Facebook 'like' or a tweet) and in order to prepare the data to enter a predictive model some pre-processing will be necessary. Often the data will need to be temporally aggregated to match the temporal aggregation level of the outcome, for example, monthly data. Also as some of the inputs from social media are text variables, some filtering, interpretation, and classification may be necessary. An example of the latter would be the application of a supervised machine learning algorithm that classifies the posts and comments into positive, negative or neutral sentiments (Thelwall, Chapter 32, this volume). At the current moment it is mainly the preprocessing of the social media data that is considered challenging from the computational aspects of big data analytics (Council, 2013). Once the individual actions (posts, likes, etc.) are temporally aggregated and classified, the set of potential explanatory factors are usually rather limited and as the outcome variables are of fairly low frequencies like monthly or quarterly (stock market data are actually sometimes used at a daily frequency) which means that the modelling process deviates less from more classical approaches within predictive modelling.

IV.4. In search of a model equation – theory-based versus data-driven?

In very general terms a model equation will identify some relationship between the phenomenon of interest (y) and a set of explanatory factors. The relationship will never be perfect either due to un-observable factors, measurement errors or other

types of errors.

The general equation: y = f(explanatory factors) + error

Where f describes some relationship between what is inside the parenthesis and y.

In principle, linear, non-linear, parametric, non-parametric and semi-parametric models may be considered. In general, non-linear models will require more data points/observations than linear models as the structures they search for are more complex.

There is a range of possible starting points for the search process. At one end lies traditional econometrics where the starting point is often an economic or behavioural theory that will guide the researcher in finding a set of potential explanatory factors. At the other end of the range machine learning algorithms will help identify a relationship from a large set of social media data and other potential explanatory factors. The advantage of starting from a theory-based model specification is that the researcher may be more confident that the model is robust in the sense that the identified relationship is reliable at least for some period of time. Without a theory the identified structure may still work for predictions in the short run but may be less robust and in general will not add much to an understanding of the phenomenon at hand. In between pure theoretically inspired models and models based on data pattern discoveries are many models that include elements of both categories. As theoretical models are often more precise when it comes to selection of explanatory factors for the more fundamental or long-run relationships they may be less precise when it comes to a description of dynamics and a combination that allows for a primary theoretically based long-run part may prove more useful.

To finalize the discussion of theory-based versus data-driven model selection the

concept of causality is often useful. If a causal relationship exists a change in an explanatory factor is known to imply a change in the outcome. A model that suffers from a lack of a causal relationship suffers from an endogeneity problem (a concept used in econometrics). A model that suffers from an endogeneity problem will not be useful for tests of a theory of for policy evaluations. If the only purpose of the model is forecasting, identification of a causal relationship is of less importance as a strong association between the explanatory factors and the outcome may be sufficient. However, without causality the predictive model may be considered less robust (more risk of a model break-down) to general changes in structures and society and hence may be best at forecasting in the short run. If this is the case, some sort of monitoring on a continuous basis to identify a model break-down at an early stage is advisable.

IV.4.1. Fitting of a predictive model

In this step the researcher will adapt the mathematical specification of the predictive model to the actual data. In the case of a linear regression model this is done by estimation using the ordinary least squares (OLS) method or the maximum likelihood (ML). For non-linear models such as neural networks, some mathematical algorithm is used. In rare cases estimation of a model is not possible (e.g. in case of perfect multicollinearity of a linear regression model). In such a case the researcher has to re-think the model specification.

Estimation (the use of a formula or a procedure) may in itself sound simple, but already at this stage the researcher has to specify the set-up to be used for model evaluation in the following step as they are highly dependent.

Even though it may seem natural to use as many data point as possible for the model fitting, there are other considerations to take into account as well. For the estimation step, it is stressed that in addition to the decision of estimation or fitting

method, a decision on exactly which sample or part of the sample to use for estimation is of importance too.

IV.4.2. Evaluation of a predictive model for forecasting purposes

The true test of a predictive model that is to be used for forecasting of future values of the outcome of interest is by investigating the out-of-sample properties of the model.

This statement calls for the need of an estimation (or training) sample and an evaluation (or test) sample. As a good in-sample model fit does not ensure good forecasting properties of a predictive model, the evaluation process then naturally starts by an analysis of the in-sample properties of the model and extends to an out-of-sample analysis.

IV.4.3. In-sample evaluation of the model

The first thing to note is that if the model has a theoretical foundation the signs of the estimated coefficients will be compared to the signs expected from the theory.

A second thing to be aware of is whether the model fulfils the underlying statistical assumptions (these may differ depending on the type of model in focus). In classical linear regression modelling, problems such as autocorrelation and heteroscedasticity will need attention and a study of potential outliers is of high importance. When forecasting is the final purpose of the model multicollinearity is of less importance. Finally, indicators in relation to the functional form specification may provide useful information on how to improve the model.

The overall fit of the model may be captured by measures such as R^2, adjusted R^2, the family for measures based on absolute or squared errors (e.g. MSE, RMSE,

MAE, MAPE), and information criteria such as AIC, and BIS. A small warning is justified here as too much emphasis on obtaining a good fit may result in overfitting of the model which is not necessarily desirable when the purpose of the model is forecasting.

IV.4.4. Out-of-sample evaluation

For an out-of-sample evaluation study the model is used to forecast values for a time period that was not used for the estimation of the model. In the 'pure' case neither future values of the explanatory factors nor future values of the outcome are known and the model that is used to obtain the forecast will need to rely on lagged values of the explanatory factors or to use predicted values of the explanatory factors. In the former case, the specification of the model equation in terms of lags will set a limit to how many periods into the future the model can predict. In many cases an out-of-sample forecast evaluation will rely on sets of one step ahead predictions, but predictions for a longer forecast horizon (e.g. six months ahead for a model specified with monthly data) are also sometimes considered.

Once the out-of-sample forecasts are obtained it is possible to calculate forecast errors and to study their patterns. Focus areas will be of directional nature (the trend in the outcome captured), as they may be related to predictability of turning points and summary measures for the errors will again prove useful (e.g. MSE, MAPE, etc.) but this time for the forecasted period only. The idea of splitting the sample into different parts for evaluation can be extended in various ways using cross-validation (Hyndman & Athanasopoulos, 2014).

IV.4.5. Using a predictive model for forecasting purposes

Once a model has been chosen some considerations concerning its implementation

are important. This topic is very much related to the overall phenomenon and problem; hence a general discussion is difficult to provide.

There is, however, one type of considerations that deserves mentioning: how often the model needs re-estimation or specification updating. Given that often the general data pattern is quite robust, the specification updating may only take place in case of new variables becoming available or in case a sufficiently large number of data points have become available such that more complex structures could be allowed for.

Finally, from a practical perspective a combination of forecasts from different basic predictive models is also a possibility and quite popular in certain fields.

IV.5. Categorized List of Predictive Models with Social Media Data

Table 20.1 below presents a selected list of research papers on predictive analytics with social media data categorized across different application domains in terms of social media platform (Facebook, Twitter, etc.) and the independent and dependent variables involved. For conceptual exposition and literature review on the predictive power of social media data (see Gayo-Avello et al. (2013)).

IV.5.1. Application Domains

As can be seen from Table 20.1, there have been many predictive models of sales based on social media data. Such predictive models work for the brands that can command large amounts of human attention on social media, and therefore generate big data on social media. Examples are iPhone sales, H&M revenues, Nike sales, etc., which are all product categories around which there is a possibility to have large volumes and ranges of opinions on social media platforms. For brands and products that don't generate large volumes of social media data, for instance, insurance, banking, shipping, basic household supplies, etc. the predictive models

tend not to work. One explanation for the successful performance of the predictive models is that social media actions can be categorized into the phases of the different domain-specific models from the application domains of marketing, finance, epidemiology, etc. For example, the actual stock price for Apple is in rough terms mainly based on discounted historical sales and expectations to future sales. If social media can model sales, then there is a high potential for the associated stock price to also being modelled with social media data. In the case of epidemiology, all social media texts on flu can also be categorized in to the different domain-specific phases of spread, incubation, immunity, resistance, susceptibility etc.

Table 20.1 Categorization of Research Publications on Predictive Analytics with Social Media Data

Reference	*Social Data*	*Dependent Variables*	*Independent Variables*	*Statistical Methods*
Asur & Huberman (2010)	Twitter	Movie revenue	Twitter activity, sentiment and theatre distribution	Time-Series Multiple Regression Model
Lassen et al. (2014)	Twitter	iPhone sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Bollen & Mao (2011)	Twitter	Dow Jones Industrial Average	Calm, Alert, Sure, Vital, Kind and Happy	Time-Series Multiple Regression Model
Voortman (2015)	Google Trends	Car sales	Google trend data car names	Time Series Linear Regression Model
Vosen & Schmidt (2011)	Google Trends	Consumer spending	Real personal income y, interest rates on 3-month Treasury Bills I and stock prices s (measured on S&P 500), Google Trend, and consumer spending t-1	ARIMA/Time Series Multiple Regression Model
Choi & Varian (2012)	Google Trends	Sales of cars, homes and travel	Historical sales and Google trend variable	Simple Seasonal AR Models and Fixed-Effects Models
Chung & Mustafaraj (2011)	Twitter	Political election outcome	Twitter collective sentiment	Linear Regression
Conover, Gonçalves, Ratkiewicz, Flammini, & Menczer (2011)	Twitter	Political alignment	Twitter hashtags	SVM trained on hashtag metadata
Bothos, Apostolou, & Mentzas (2010)	IMDB, Flixster, Yahoo Movies, HSX, Twitter, RottenTomatoes.com	Movie Academy Award winners	Measures from IMDB, Flixster, Yahoomovies, HSX, Twitter, RottenTomatoes.com	Multivariate Distribution Models
Culotta (2010)	Twitter	Detecting influenza outbreaks	Twitter keywords	Time-Series Multiple Regression Model
Dijkman, Ipeirotis, Aertsen, & van Helden (2015)	Twitter	Many types of sales	Twitter activity and sentiment	Time-Series Multiple Regression Model
Eysenbach (2011)	Twitter	Total number of citations	Twimpact variable (number of tweetations within n days after publication)	Multi-Variate/Linear Regression
Gruhl, Guha, Kumar, Novak, & Tomkins (2005)	Blogs	Sales	Product/brand mentions	Time-Series using Cross-Correlation
Jansen, Zhang, Sobel, & Chowdury (2009)	Twitter	Brand variables	Twitter sentiment variables	Time-Series Linear Regression Models
Li & Cardie (2013)	Twitter	Early stage influenza detection	Twitter texts about flu	Unsupervised Bayesian Model based on Markov Network
Radosavljevic, Grbovic, Djuric, & Bhamidipati (2014)	Tumblr	Sport results and number of goals	Team and player mentions	Poisson Regression Model using Maximum Likelihood Principle
Ritterman, Osborne, & Klein (2009)	Twitter	Stock-Prices	Historical prices, unigrams and bigrams, Twitter activity	SVR Regression using Unigrams and Bigrams

Sang & Bos (2012)	Twitter	Dutch election outcome	Twitter texts and sentiments	Time Series Multiple Regression Model
Shen, Wang, Luo, & Wang (2013)	Twitter	Entity belonging	Twitter texts	Decision tree, KAURI/LINDEN method
Skoric et al. (2012)	Twitter	Election outcome Singapore	Twitter activity	Time Series Linear Regression Model
Tsakalidis et al. (2015)	Twitter	Election outcomes EU	Twitter texts	Linear Regression (LR), Gaussian Process (GP) and Sequential Minimal Optimization for Regression (SMO)
Tumasjan, Sprenger, Sandner, & Welpe (2010)	Twitter	Election outcomes Germany	Twitter texts	Probability Models
Yu, Duan, & Cao (2013)	Google blogs, Boardreader and Twitter compared to Google News	Firm equity value	Variables for activity and sentiment	Time Series Multiple Regression Model
Hughes, Rowe, Batey, & Lee (2012)	Twitter and Facebook	Socialising and info exchange	Big5 personality traits, NFC and sociability	Time-Series Multiple Regression Model
Krauss, Nann, Simon, Gloor, & Fischbach (2008)	Forums	Movie success and academy awards	Intensity, positivity and trendsetter variables	Time-Series Multiple Regression Model
Seiffertt & Wunsch (2008)	Several	Variables on financial markets	Many types discussed	Different Model Types Discussed
Tang & Liu (2010)	Flickr and YouTube	Online behaviors	Social Dimension variables	SocioDim, several advanced models combined
Karabulut (2013)	Facebook	Stock prices	Facebook GNH (General national happiness), positivity, negativity	Time-Series Multiple Regression Model
Mao, Counts, & Bollen (2014)	Twitter	UK, US, and Canadian stock markets	"Bullish" or "bearish" mentions on Twitter	Time-Series Multiple Regression Model
Bollen, Mao, & Zeng (2011)	Twitter	DJIA	Twitter moods and feelings	Self-Organizing Fuzzy Neural Network
Bollen, Mao, & Pepe (2011)	Twitter	Socio-economic events	Twitter moods and feelings	Extended version of: Profile of mood states
Eichstaedt et al. (2015)	Twitter	Heart attacks	Anger, stress and fatigue	Time-Series Multiple Regression Model
De Choudhury, Gamon, Counts, & Horvitz (2013)	Twitter	Depression	Language, emotion, style, ego-network, and user engagement	Support Vector Machine
De Choudhury, Counts, & Horvitz (2013)	Twitter	Postpartum changes in emotion and behaviour	Engagement, emotion, ego-network and linguistic style	Support Vector Machine
De Choudhury, Counts, Horvitz, & Hoff (2014)	Facebook	Postpartum depression	Social activity and interaction	OLS Regression Model
Weeks & Holbert (2013)	Facebook, Twitter, YouTube	Dissemination of News Content in Social Media	Gender, age, web engine news search, email news activity and cell phone activity	Decision Tree Model
Gilbert & Karahalios (2009)	Facebook	Tie strength	15 communication variables	Time-Series Multiple Regression Model
Won et al. (2013)	Weblog social media data	Suicide	Suicide related words and mentions	Time-Series Multiple Regression Model

IV.5.2. Social Media Data Types

For modelling stock prices, Twitter and Google Trends have proven to be the best platforms. Twitter and Google Trends beat Facebook for stock price modelling because of higher data volume and immediacy. On the other hand, Facebook data have been successfully used for modelling sales, human emotions, personalities and human relations to a brand. In general, picture and video based social media platforms such as Instagram, YouTube and Netflix are becoming more prevalent and we expect them to become more relevant for predictive models in the future.

IV.5.3. Independent and Dependent Variables

As can be seen from Table 20.1, a wide range of dependent variables have been modelled: sales, stock prices, Net Promoter Score, happiness, feelings, personalities, interest areas, social groups, diseases, epidemics, suicide, crime, radicalization, civil unrest. The independent variables used reflect the human social relations to the dependent variables mainly consist of measures of social media activity, feelings, personalities and sentiment.

IV.5.4. Statistical Methods Employed

We find that a wide range of statistical models for predictive analytics have been used including Regression, Neural Network, SVM, Decision Trees, ARIMA, Dynamic Systems, Bayesian Networks, and combined models.

In the next section, we present an illustrative case study of predictive modelling with big social data.

IV.6. An Illustrative Case Study of Predictive Modelling

In this section, we demonstrate how social media data from Twitter and Facebook can be used to predict the quarterly sales of iPhones and revenues of clothing

retailer, H&M, respectively. Based on a conceptual model of social data (Vatrapu, Mukkamala, & Hussain, 2014) consisting of Interactions (actors, actions, activities, and artifacts) and Conversations (topics, keywords, pronouns, and sentiments), and drawing from the domain-specific theories in advertising and sales from marketing (Belch, Belch, Kerr, & Powell, 2008), we developed and evaluated linear regression models that transform (a) iPhone tweets into a prediction of the quarterly iPhone sales with an average error close to the established prediction models from investment banks (Lassen et al., 2014) and

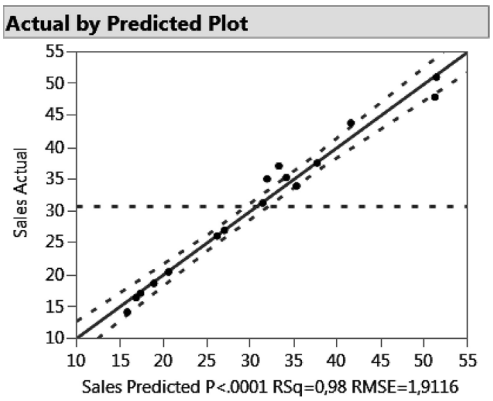
(b) Facebook likes into a prediction of the global revenue of the fast fashion company, H&M. Our basic premise is that social media actions can serve as proxies for user's attention and as such have predictive power. The central research question for this demonstrative case study was: *To what extent can Big Social Data predict real-world outcomes such as sales and revenues?* Table 20.2 below presents the dataset collected for predictive analytics purposes of this case study. We adhered to the methodological schematic recommended by Shmueli & Koppius (2011) for building empirical predictive models. We built on and extended the predictive analytics method of Asur & Huberman (2010) and examined if the principles for predicting movie revenue with Twitter data can also be used to predict iPhone sales and

Table 20.2 Overview of Dataset

Company	*Data Source*	*Time Period*	*Size of Dataset*
Apple	Twitter	01–2007 to 10–2014	~500 million+ tweets containing "iPhone" Collected using Topsy Pro (http://topsy.thisisthebrigade.com)
H&M	Facebook	01–2009 to 10–2014	~15 million data points from the official H&M Facebook page Collected using the Social Data Analytics Tool (Hussain & Vatrapu, 2014)

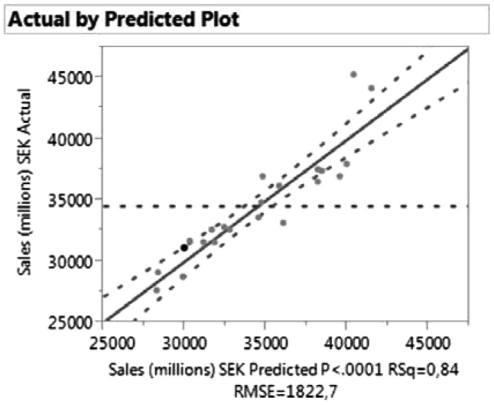
H&M revenues for Facebook data. That is, if a tweet/like can serve as a proxy for a user's attention towards a product and an underlying intention to purchase and/or

recommend it. We extend Asur & Huberman (2010) in three important ways: (a) addition of Facebook social data, (b) theoretically informed time lagging of the independent variable, social media actions, and (c) domain-specific seasonal weighting of the dependent variable, sales/revenues. Figures 20.1 and 20.2 present the predicted vs. actual charts for Apple iPhone sales and H&M revenues respectively. With regard to our prediction models, we observed a 5–10% average error from our predictive models with the actual sales and revenue data over three-year period of 2012–2014. In the case of the iPhone sales prediction model, our average error of 5% is not that far from the industry benchmark predictions of Morgan Stanley and IDC. That said, there are several challenges and limitations to the predictive analytics processes and their outcomes. First, we lack multiple cases to extensively evaluate and validate the overall prediction model. A second limitation is the emerging challenge for predictive analytics from social data associated with increasing sales in emerging markets such as China with its own unique social media ecosystem. By and large, the social media ecosystem of China does not overlap with that of Western countries to which Facebook and Twitter belong. We suspect that the effect of



Summary of Fit	
RSquare	0,975862
RSquare Adj	0,972149
Root Mean Square Error	1,91163
Mean of Response	30,72688
Observations (or Sum Wgts)	16

Figure 20.1: Predictive Model of iPhone sales from twitter data



Summary of Fit	
RSquare	0,841863
RSquare Adj	0,833956
Root Mean Square Error	1822,675
Mean of Response	34508,05
Observations (or Sum Wgts)	22

Figure 20.2: Predictive Model of h&M revenues from facebook data

non-overlapping social media ecosystems might be somewhat ameliorated for Veblen goods such as iPhones given the conspicuous consumption aspirations of a global middle class. This however remains an analytical challenge and restricts the predictive power of our H&M prediction model.

IV.7. Conclusion

Predictive models offer powerful tools as numerical forecasts and assessments of their uncertainty alongside quantitative statements more generally may improve decisions in companies and by public authorities.

The overall advice is to go for a parsimonious, simple model that captures the most important features of the data, that fulfils the model assumptions and that provides a good fit both in sample and out of sample. Furthermore, it is important that even during the phase where the model is applied for its purpose, it performance is still monitored. We present a general model for predictive analytics of business outcomes from social media data below.

$$yt = \beta a \times At + \beta p \times Pt + \beta d \times Dt + \beta o \times Ot + \varepsilon t$$

Where:

y_t = Outcome variable of interest

A_t = Accumulated time-lagged social media activity associated with outcome variable at time t set of information.

$A_t = \Sigma A_{st}$

A_{st} = Social media activity in terms of actions by actors on artifacts associated with outcome variable at time t

P_t = Individual or social psychological attribute(s) at time t

D_t = Social media dissemination factors

O_t = Other explanatory factors

A final word of caution will end this chapter: any predictive model is based on a certain set of information. It is necessarily backward-looking as it relies on historical data and irrespectively of how carefully the model specification and evaluation is done, there is no guarantee that the prediction of future values of the variable of interest will be reliable. The patterns or theories that the model relies on may break down and render the model useless for predictive purposes. That being said, careful predictive modelling is probably the best that can be done and, if applied and used following the state of the art with most emphasis placed on short term forecasting, predictive modelling is a very valuable tool.

Acknowledgements

We thank the members of the Centre for Business Data Analytics (http://bda.cbs.dk) for their feedback.

The authors were partially supported by the project Big Social Data Analytics: Branding Algorithms, Predictive Models, and Dashboards funded by Industriens Fond (The Danish Industry Foundation). Any opinions, findings, interpretations, conclusions or recommendations expressed in this chapter are those of its authors and do not represent the views of the Industriens Fond (The Danish Industry Foundation).

IV. References

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

Belch, G. E., Belch, M. A., Kerr, G. F., & Powell, I. (2008). *Advertising and promotion: An integrated marketing communications perspective*: McGraw-Hill, London.

Bollen, J., & Mao, H. (2011). Twitter mood as a stock market predictor. *Computer*, 91–94.

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM, 11*, 450–453.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Bothos, E., Apostolou, D., & Mentzas, G. (2010). Using Social Media to Predict Future Events with Agent-Based Markets. *IEEE Intelligent Systems*, *25*(6), 50–58.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*(1), 1–20.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45. doi: 10.4269/ ajtmh.2012.11–0597

Chung, J. E., & Mustafaraj, E. (2011). *Can collective sentiment expressed on Twitter predict political elections?* Paper presented at the AAAI.

Cioffi-Revilla, C. (2013). *Introduction to Computational Social Science: Principles and Applications*: Springer Science & Business Media.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). *Predicting the political alignment of Twitter users.* Paper presented at the Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE 3rd International Conference on Social Computing (SocialCom).

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Loreto, V.,

Moat, S., Nadal, J.P., Sanchez, A., Nowak, A & Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal*, *214*(1), 325–346.

Council, N. (2013). Frontiers in massive data analysis: The National Academies Press Washington, DC.

Culotta, A. (2010). *Towards detecting influenza epidemics by analyzing Twitter messages.* Paper presented at the Proceedings of the first workshop on social

media analytics.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). *Predicting postpartum changes in emotion and behavior via social media.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). *Characterizing and predicting postpartum depression from shared Facebook data.* Paper presented at the Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting Depression via Social Media.* Paper presented at the ICWSM.

Dijkman, R., Ipeirotis, P., Aertsen, F., & van Helden, R. (2015). Using Twitter to predict sales: a case study. *arXiv preprint arXiv:1503.04599*. Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap,

M. and Weeg, C. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159–169.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet Research*, *13*(4), e123.

Evangelos K, Efthimios T and Konstantinos T. (2013) Understanding the predictive power of social media. *Internet Research*, *23*(5), 544–559.

Gilbert, E., & Karahalios, K. (2009). *Predicting tie strength with social media.*

Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881.

Gruhl, D., Guha, R., Kumar, R., Novak, J., & Tomkins, A. (2005). *The predictive power of online chatter.* Paper presented at the Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery in data mining.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2), 561–569.

Hussain, A., & Vatrapu, R. (2014). Social Data Analytics Tool (SODATO). In M. Tremblay, D. VanderMeer, M. Rothenberger, A. Gupta, & V. Yoon (Eds.), *Advancing the Impact of Design Science: Moving from Theory to Practice* (Vol. 8463, pp. 368–372): Springer International Publishing, Switzerland.

Hussain, A., Vatrapu, R., Hardt, D., & Jaffari, Z. (2014). Social Data Analytics Tool: A Demonstrative Case Study of Methodology and Software. In M. Cantijoch, R. Gibson, & S. Ward (Eds.), *Analyzing Social Media Data and Web Networks* (pp. 99–118): Palgrave Macmillan, UK..

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts: https://www.otexts.org/fpp/

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188.

Karabulut, Y. (2013). *Can Facebook predict stock market activity?* Paper presented at the AFA 2013 San Diego Meetings Paper.

Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008). *Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis.* Paper presented at the ECIS.

Lassen, N., Madsen, R., & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. *Proceedings of IEEE 18th International Enterprise Distributed Object Computing Conference (EDOC 2014), Ulm, Germany*, 81–90, ISBN: 1541–7719/1514, doi: 1510.1109/ EDOC.2014.1520.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. and Jebara, T. Computational Social Science. *Science*, *323*(5915), 721–723. doi: 10.1126/science.1167742

Li, J., & Cardie, C. (2013). Early stage influenza detection from Twitter. *arXiv preprint arXiv:1309.7340*.

Mao, H., Counts, S., & Bollen, J. (2014). *Quantifying the effects of online bullishness on international financial markets.* Paper presented at the ECB Workshop on Using Big Data for Forecasting and Statistics, Frankfurt, Germany.

Radosavljevic, V., Grbovic, M., Djuric, N., & Bhamidipati, N. (2014). *Large-scale World Cup 2014 outcome prediction based on Tumblr posts.* Paper presented at the KDD Workshop on Large-Scale Sports Analytics , New York.

Ritterman, J., Osborne, M., & Klein, E. (2009). *Using prediction markets and Twitter to predict a swine flu pandemic.* Paper presented at the 1st

international workshop on mining social media, Sevilla, Spain.

Robertson, S., & Vatrapu, R. (2010). Digital Government. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 44, pp. 317–364).

Robertson, S., Vatrapu, R., & Medina, R. (2010a). Off the wall political discourse: Facebook use in the 2008 US Presidential election. *Information Polity*, *15*(1), 11–31.

Robertson, S., Vatrapu, R., & Medina, R. (2010b). Online Video "Friends" Social Networking: Overlapping Online Public Spheres in the 2008 U.S. Presidential Election. *Journal of Information Technology & Politics, 7*(2–3), 182–201. doi:10.1080/19331681003753420

Sang, E. T. K., & Bos, J. (2012). *Predicting the 2011 Dutch senate election results with Twitter*. Paper presented at the Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, France.

Seiffertt, J., & Wunsch, D. (2008). Intelligence in Markets: Asset Pricing, Mechanism Design, and Natural Computation [Technology Review]. *Computational Intelligence Magazine, IEEE*, *3*(4), 27–30.

Shen, W., Wang, J., Luo, P., & Wang, M. (2013). *Linking named entities in tweets with knowledge base via user interest modeling*. Paper presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3) 289–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems

research. *MIS Quarterly*, *35*(3), 553–572.

Skoric, M., Poor, N., Achananuparp, P., Lim, E.P., & Jiang, J. (2012). *Tweets and votes: A study of the 2011 Singapore general election.* Paper presented at the System Science (HICSS), 45th Hawaii International Conference on System Sciences, Hawaii.

Tang, L., & Liu, H. (2010). Toward predicting collective behavior via social dimension extraction. *Intelligent Systems, IEEE*, *25*(4), 19–25.

Tsakalidis, A., Papadopoulos, S., Cristea, A. I., & Kompatsiaris, Y. (2015). Predicting elections for multiple countries using Twitter and polls. *Intelligent Systems, IEEE*, *30*(2), 10–17.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM, 10*, 178–185.

Vatrapu, R., Mukkamala, R., & Hussain, A. (2014). *A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings.* Paper presented at the Computational Social Science Workshop at the European Conference on Complex Systems 2014, Lucca.

Vatrapu, R., Robertson, S., & Dissanayake, W. (2008). Are Political Weblogs Public Spheres or Partisan Spheres? *International Reports on Socio-Informatics*, *5*(1), 7–26.

Vatrapu, R., Hussain, A., Lassen, N. B., Mukkamala, R., Flesch, B., & Madsen, R. (2015). Social set analysis: four demonstrative case studies. *Proceedings of the 2015 International Conference on Social Media & Society*. doi:10.1145/2789187.2789203

Voortman, M. (2015). Validity and reliability of web search based predictions for car sales.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565–578.

Weeks, B. E., & Holbert, R. L. (2013). Predicting dissemination of news content in social media a focus on reception, friending, and partisanship. *Journalism & Mass Communication Quarterly*, *90*(2), 212–232.

Won, H.-H., Myung, W., Song, G.-Y., Lee, W.-H., Kim, J.-W., Carroll, B. J., & Kim, D. K. (2013). Predicting national suicide numbers with social media data. *PloS one*, *8*(4), e61809.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919–926.

Paper V: Google searches linked to Apple stock volatility ups and downs

Published in the Conference Book in the Proceedings of the 43rd Symposium i Anvendt Statistik, Denmark, Axelborg, Copenhagen, Denmark, presented 29. August 2022 at the conference (single-author paper).

Google searches linked to Apple stock volatility

Niels Buus Lassen[a]

[a] *Centre for Business Data Analytics, Copenhagen Business School, Denmark*

Abstract

The recent studies on social media that link news data to volatility show a Twitter buzz up is typically linked to higher volatility, while a general news media buzz is linked to lower volatility in the following month. This article demonstrates that Google searches influence Apple stock volatility in either on a weekly basis by analyzing the behavior of private and professional investors in relation to Google searches and how this behavior links to Applestock volatility. To this end, this study employs the logic of sales modeling and, thus, contributes to the theoretical construction of the novel "investor journey model" by mapping Google searches onto investorbehavior, which is an under-researched field in the literature. Subsequently, the paper summarizes the main findings in this field and outlines future challenges in this research.

Keywords: Investor behavior, Google searches, Stock markets,Investor sophistication, Decision making

V.1. Introduction

Twitter data have been included in several models and shown to have predictive powerfor both stock price indexes and specific stock price movements (see, e.g., Bollen & Mao 2010; Jiao, Veiga, & Walther 2016; Li, van Dalen, & van Rees 2018.).

A logical explanation for the predictive power of social media data in terms of financial market behavior is the size of the big data from social media chats and also Google searches about the respective stocks. Stocks with large amount of social media data, are popular stocks people like to talk about on social media and also do Google searches about.

Text mining can identify patterns in the big data from the social media and Google searches. Statistical and machine learning methods can be tested to model the human behavior on social media and Google searches to financial market behaviors.

This approach is comparable to the sales of products and services, where big data from social media can be used to predict sales. The sales models that build on social media data work well if social media data are big enough or, inother words, if customers like to talk about a product or service on social media. Examples are H&M, Nike, and Apple (see, e.g., Lassen et al. 2017; Boldt et al. 2016). These are popular social media topics that produce big enough data to have predictive power for their sales.

All chat text on social media about products and services can be categorized into one of the phases in the Customer Infinity Model (Figure 1). These phases of customer behavior can then be modeled based on sales and provide logical explanation for why social media data can predict sales, if the data are big enough

(see, e.g., Asur & Huberman 2010; Lassen et al. 2014).

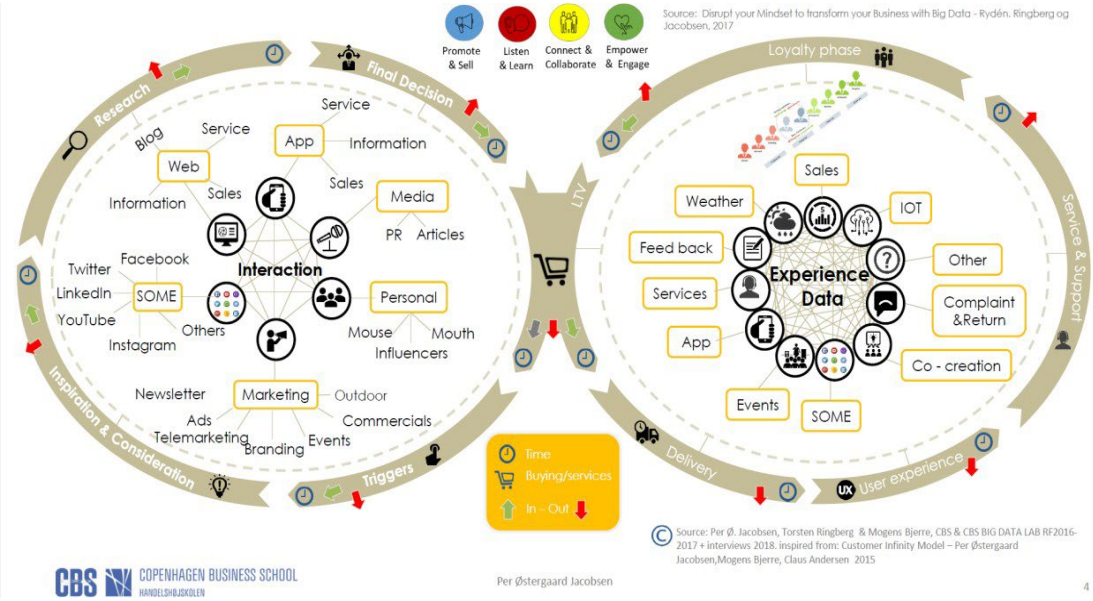


Fig. 1: Customer Infinity Model

The logic for stock prices is comparable, based on which this article proposed the "investor journey model," which posits that all social media text or web searches can be categorized into one of the phases of this model, which can in turn be linked to Apple stock volatility, as well as other stocks with big enough data on social media and from web searches. This paper focuses on both the private and professional investor behavior related to Google searches.

More than 90% of all global web searches use Google; specifically, approximately 63% on the search engine, 23% on Google Images, 4% onYouTube, and 1% on Google Maps (BusinessInsider.com 2018). When amateur and professional investors search for information about the stocksthey are interested in, Google searches provide relevant stock information and form the majority of all web searches. The model proposed in this article has been tested on more than 60 Google searches before identifying good proxies for the amateur and professional investor

behavior based on Google searches for Apple stocks.

The research questions this article tackles as follows:

RQ1: Can *Google search data predict stock price volatility?*

RQ2: Which *Google searches are creating ups and downs inApple stock price volatility?*

RQ3: *Can* the *identified ups and downs in Apple stock pricevolatility be linked to household and professional investor activity?*

The contribution of this article is identifying new patterns for investor behavior on web searches, and how these patterns link to ups and downs instock volatility theoretically explained through the proposed investor journey model, which relies on insights from related queries and topics in Google Trends from more than 60 Apple-related Google searches. Apple is the example used in this article, and model may be applicable for similar big tech stocks with high volume google search data.

V.2. Literature review

One of the most notable articles modeling stock price volatility using Google Trends, is Preis et al. (2013), which mentions: "We suggest that *Google Trends* data and stock market data may reflect two subsequent stages in the decision making process of investors."

This article suggests that Apple stock-related Google Trends data follow an investor decision making journey, which affects Apple stock volatility. The proposed investor journey model is detailed in section 3.

Jiao, Veiga, and Walther (2016) find that a buzz up in coverage by traditional news media predicts subsequent decreases in volatility and turnover, while a buzz up in

coverage by Twitter predicts increases in the subsequent volatility and turnover. However, they do not explain why the buzz in traditional newsmedia and Twitter coverage have these different effects on stock volatility.

Greenwich Associates published a report in 2015 based on interviews with 256 asset owners from more than 250 institutional investor organizations, which shows that institutional investors use Twitter in a very limited manner in their decision-making process, compared to LinkedIn, for example. Among the 256 interviewed institutional investors, LinkedIn was used by more than half and often played an important role in investor decision making. The interviewed institutional investors recognized the value of the Twitter news feed in seeking opinions or commentary on market events, but considered LinkedIn feeds to be better targeted, as they reflected their professional ties more closely.

This article does not include LinkedIn data, but only Apple-related Googlesearches due to their free availability through Google Trends.

Specifically, LinkedIn data are difficult to access and expensive. Twitter data were also not considered due to their cost for this article.

Institutional and professional investors use information processed from Twitter by analytical companies such as Dataminr. For example, TheGlobeAndMail.com (2018) states: “Dr. Mohanram cautions that individual investors are not likely to be able to correctly replicate the conditions of the study to benefit from crowd-sourced opinion. It requires a certain amount of data crunching ability and sophistication [to] analyze a large sample of tweets, categorize them and aggregate them all in real time.”

Private investors are using more Twitter unprocessed information in their decision-making processes such as by following investor gurus and searching for stock

related information on Twitter. For example, MarketWatch.com's (2018) "Finance Twitter: The 50 most importantpeople for investors to follow" is one of many articles recommending private investors who to follow on Twitter to get smart insights on investing. Based on such recommendations, it becomes logical that the abundance of financial gurus on Twitter is creating noise and increases stock volatility. Searching for stock related information on Twitter typically yields several credible news sources along with sources that require vetting. In short, the many news sources on Twitter create noise and increase stock volatility. Namely, private investors are more exposed to the risk of rumors, old news being perceived as new news, speculations, and manipulating info to drive stock prices up or down. All this noise can lead to irrational investor behavior by creating higher volatility when there is a buzz up in Twitter info for a stock.

Therefore, the logic of Twitter usage is that private investors are the main actors in the increased stock volatility after a buzz up in the Twitter infoabout a stock.

Traditional news media info about stocks are typically used by professional and institutional investors. Examples of reliable traditional news media channels are Financial Times, Bloomberg, Wall Street Journal, Reuters, CNBC, Forbes, and MarketWatch. They are associated with rational investor behavior and lower volatility after a news buzz up related to a specific stock.

For the Apple stock, around 60–70% of Apple stock trading is conducted by professional and institutional investors. For details, please refer to https://finance.yahoo.com/quote/AAPL/holders/.

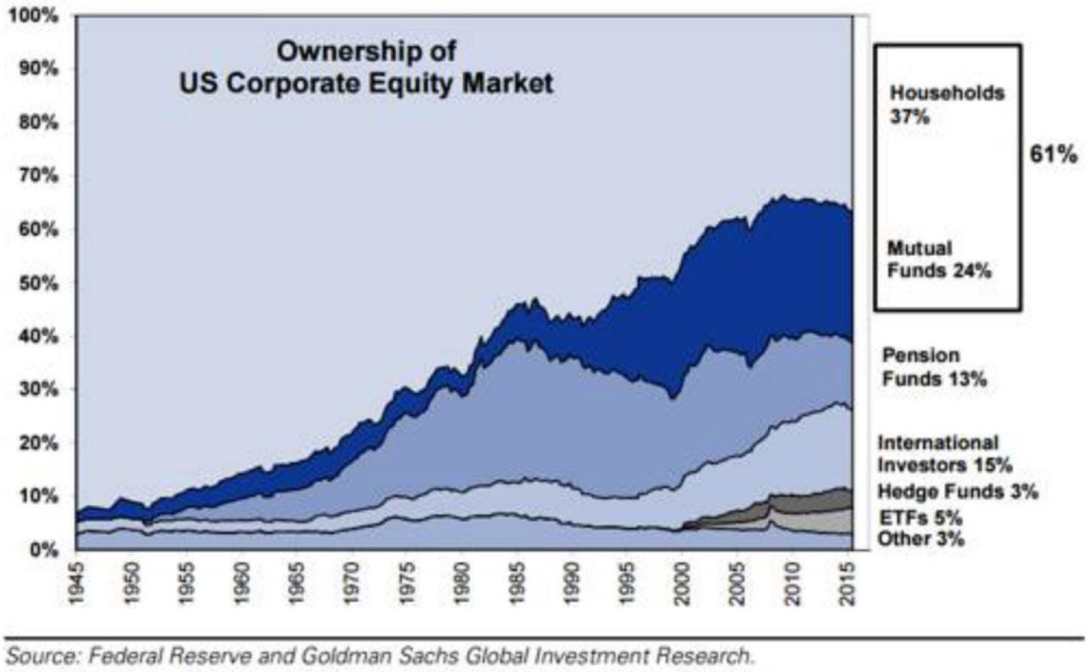


Fig. 2: Ownership of the US corporate equity market: USD 36 trillion as of Q2 2016 and includes USD 7 trillion of foreign equity holdings.

Based on Fig. 2, US households own approximately one third of the US corporate equity market, meaning that the household investor activity in Apple stocks is estimated to be one third on average.

Many models for stock investor decision making are based on the prospect theory developed by Kahneman and Tversky (1979), which is a descriptive model of decision making under risk (see, e.g., Wakker & Levy 2015). Prospect theory has been criticized and several alternatives have been suggested (see, e.g., Nwogugu 2005; Levy et al. 2002). Prospect theory assumes decision making has two phases: editing and valuation. The editing phase refers to investors scanning the information and forming their beliefs on eventual outcomes. This phase is heavily dependent on psychological biases. The second phase is the valuation phase, where agents value these eventual outcomes based on the beliefs in the first phase. This phase depends more on the risk preferences of investors. These two decision-

making phases are distinct, albeit closely dependent on each other.

The model for stock investor decision making developed in this article isbased on the empirical observation of Google searches related to the Apple stock (stock symbol: AAPL). The model does contain some editing and valuation by investors but focuses on explaining how different types of investor use information before, during, and after the Apple Quarterly Reports and iPhone models are released.

V.3. The Investor Journey Model

The "investor journey model" is a conceptual model developed in this study for the analysis of investor behavior based on web searches and is applied to Apple stocks in this paper. Specifically, it has been developed based on related queries and topics on Google Trends from more than 60 Apple-related Google searches. It is thus the main analysis framework in this study, as described in Figs. 3 and 4.

		Household investors	Professional & Institutional investors
	Phase 1	Would I like to buy or sell the Apple stock? Or should I buy or sell Google, Amazon, Facebook, Microsoft, Nvidia or IBM instead?	Should I add or decrease Apple stocks in my portfolio? The Apple stock is here in a setup of portfolio diversification & risc profiles.
	Phase 2	Which stock or stocks do I proceed with? Which rumor news will affect this?	What is my analysis of Apple sales & results, before the Apple Quarterly Report? Any reliable Apple rumor news indicating level of sales before the Apple Quarterly Report?
Every end of Jan, Apr, July & Oct	Phase 3a	Apple Quarterly Report release. Above or under my expectations? Time to buy or sell Apple stock?	Apple Quarterly Report release. Above or under my analysis? Should I add or decrease Apple stocks in my portfolio?
Every Sep.	Phase 3b.	Would I like to buy or sell the Apple stock? before the Sep launch of new iPhone? Will the iPhone launch drive stock up or down?	What is my analysis of iPhone launch, before the Sep launch of new iPhone? What is the level of rumor news compared to earlier years? Will the iPhone launch drive stock up or down?

Fig. 3: The Investor Journey

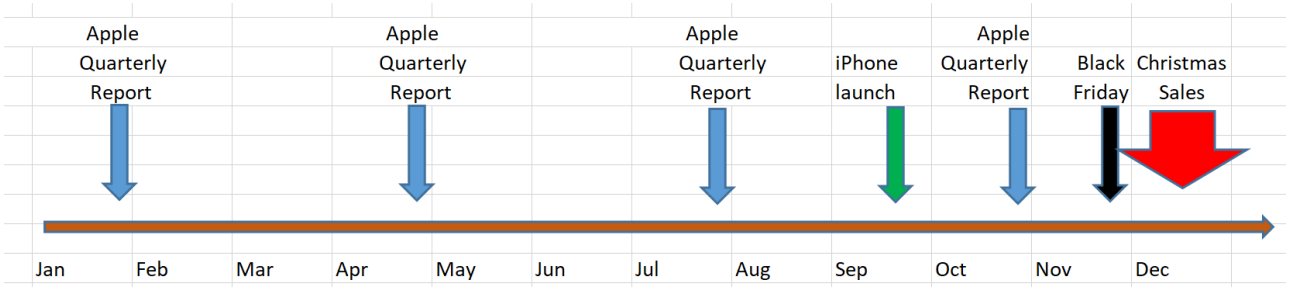


Fig. 4: Timeline for the investor journey

I divide the Apple stock-related Google searches into two groups—household and professional investors—based on the following hypotheses.

Hypothesis 1: The Google searches for stock symbols AAPL, AMZN, and IBM are assumed to be mostly made by household investors, as professional investors are likely familiar these stock symbols and do not need to Google search them. Google searches for these stock symbols lead to SeekingAlpha.com and StockTwits.com, which have large groups of household investors among their readers.

Hypothesis 2: The Google searches for Apple rumor news are assumed to be largely done by professional investors, as the Apple rumor news sites use high vetting levels and due diligence for the sources.

V.4. Methodology

The analysis is based on the multiple regression modeling of more than 60Apple-related Google searches as input variables and investor behavior inthe form of Apple stock volatility as the dependent variable.

For the Apple and rumor news Google searches, professional investors largely consider the vetted news and rumor sites, while household investors largely Google the stock symbols of Appleand other big tech companies. These patterns were identified based on the related topics and queries for each Google search, but also on the assumption of the tendencythat household investors drive volatility up and professional investors drive it down. This argumentation of the patterns for

household and professional investors will be further elaborated upon in the following sections.

The statistical software used is Oxmetrics 8.10, which mainly focuses on using the Autometrics functionality. Autometricsis a part of the PcGive module of Oxmetrics, being the automatic econometric model selection procedure that is available in PcGive. It is based on regression modeling under the general unrestricted model (GUM) framework. In Autometrics, the variable and model selection criteria are based upon the unique method developed by David Hendry and Jürgen Doornik, which performs well on gauge and potency. Gauge is the retention rate of irrelevant variables in the selected model and akin to size, because it accounts for the wrongly selected variables . Potency is the retention rate of the relevant variables in the selected model. It is also akin to size because it accounts for the variables that have been correctly selected (see Hendry et al. 2014).

The chosen target size for the dataset used in this study is 1%, which is the t-probability threshold for choosing and eliminating input variables. The1% level was chosen because of the 60+ Google searches selected as input variables in 1–4 time lags each; therefore, the large amount of input variables could be cut down to a reasonable number. After selecting among more than 60 Google searches with a target size of 1%, a target size of 5% was also tested when the input variables were reduced to a group of 10 predictors.

Another variable selection method that comes from machine learningis the least absolute shrinkage and selection operator (LASSO) method. It is a type of linear regression that uses shrinkage, where data values are shrunk towards a central point, such as the mean. LASSO adds the “absolute value of magnitude” of the coefficient as penalty term to the loss function, which will be minimized. This is

called L1 regularization.

This regression works especially well for many input variables and multicollinearity and can limit the input variables significantly. The input variables field is cut down by the described L1 regularization.

The analysis will examine if Apple-related Google searches are good proxies for investor behavior regarding the Apple stock. Both household and professional investor are reflected in Google searches on the investorjourney to buy or sell Apple stocks.

V.5. Data

V.5.1. Google searches during 2015–2020

The Google search data were collected from Google Trends (https://trends.google.com/trends/). Specifically weekly Google search data were collected from April 2015 to April 2020, as this is the longest period available on Google Trends with weekly Google search data.

It is possible to get obtain Google searches on Google Trends, but they are only available up to the 90 prior days. Weekly Google search data are available on Google Trends for the past 12 months or 5 years. For longer periods, starting from 2004, monthly date are available. The selected data were evaluated to be the most suitable dataset for modeling Applestock prices and volatility because it was the longest and most recent time period available with weekly data at the time of this study.

The Google search data are given in indexes from 0 to 100, as positive integers, available on Google Tends. An index 100 for one or more weeks would be the highest weekly search volume for the entire 5-year period. In this article, more than 60 Apple-related Google searches were extracted for the 5-year period and tested

for their relationship to stock price and volatility. These searches were found by exploring Google Trends for the Apple stock symbol (AAPL) and products such as iPhone, iPad, MacBook iOS, or MacOS. Google Trends includes both related topics and queries, based on which I found stocks related to the Apple stock, which led to the idea of developing the investor journey model. Fig. 5 shows one data extract.

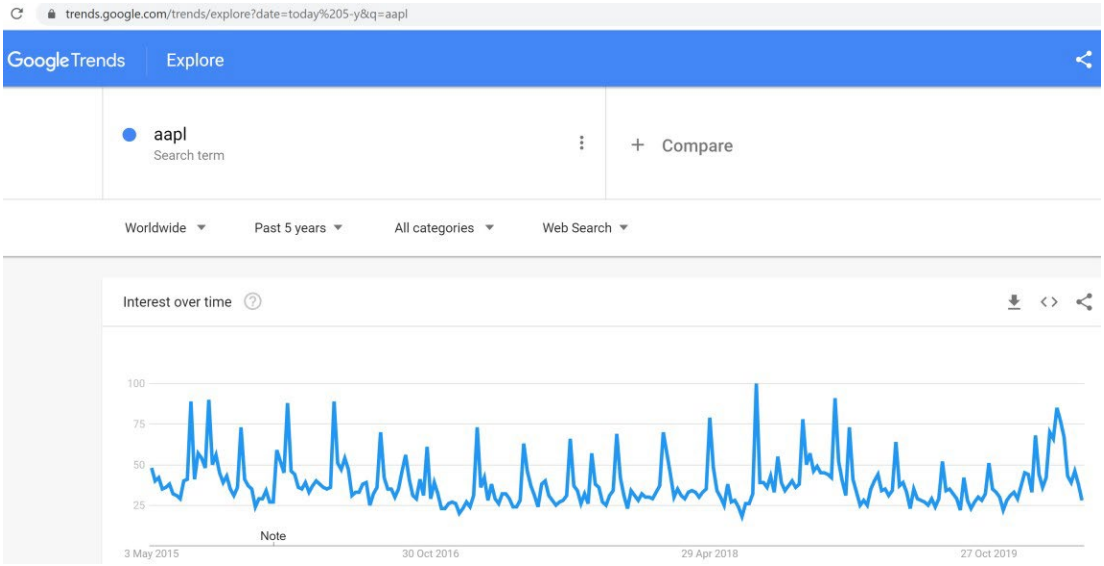


Fig. 5: Google search data for "AAPL" for the period 2015-2020

Source: https://trends.google.com/trends/explore?date=today%205-y&q=AAPL

The Google searches can be extracted in sets of up to five, but I instead considered one search at a time, as for more than five searches extracting sets will create problems. That is, because the five searches will be indexed in a group of five, a second set will not be indexed against the first set. As such, unless there is an overall baseline, the highest of all Google searches will be identified for each set. However, the baseline can change during the research to include all the dataset. As such, the most practical approach is to extract Googlesearches one at a time and, in the end, the searches can be indexed together in one set of five or two sets of 10 for 5–10 input variables.

For example, for the Google search of the Apple stock symbol (AAPL) shownin

Fig. 5, all weeks in the 5-year period are indexed around the datapoint with index 100 in week 31 of 2018. For up to 5 searches in the same Google Trends query, all weeks would be indexed around the highest search with index 100 in a given week. The final dataset fromGoogle Trends included sets of five Google searches and all searches were indexed to the highest search index, "IBM," in the 5-year period.

Variables	*N*	Mean	Median	St. dev.	Min	Max
AvgWeeklyClose	260 weeks	164.5	158.1	54.3	91.9	323.6
Weekly Volume, number of million shares	260 weeks	170.8	156.4	72.7	32.5	500.4
Weekly volatility	260 weeks	3.20%	2.64%	2.24%	0.51%	19.17%
First diff Log(avgWeekClose)	260 weeks	0.13%	0.24%	1.69%	-13.46%	12.76%

Table 1: List of weekly financial variables

Table 1 shows that the weekly volatility includes an extreme outlier of 19.17%. This was due to the COVID-19 pandemic from February 24 to April 12, 2020, were the weekly volatility ranged between 9% and 19.2%, peaking at 19.2% in week 11—March 9–15, 2020. In the same time window, the average trading volume was 300 million shares per week and stock price varied from USD 212 to USD 304.

Table 2 presents the descriptive statistics for the Google search data. As previously mentioned, all Google searches for 5 years have been downloaded from Google Trends as 260 weekly observations. Specifically, 62 Apple related Google searches were tested for modeling the Apple stock volatility, among which nine were significant and were chosen for further modeling. The two most important input variables, the Google searches for **"AAPL"** and **"AMZN"**, selected by SPSS

LASSO and Autometrics are marked with **bold.** The Google searches for **"MacRumors"** and **"Apple rumors"** are also marked with **bold**, as they were additionally selected by Autometrics when the target size was changed from 1% to 5%. The target size in Autometrics is the t-probability threshold for choosing and eliminating input variables. There are two groups for the searches:

1. News and rumors searching/vetting for Apple stock
2. Apple and other related big tech stock searches

The Google search "Apple rumors" marked with green is the only variable considered with an overweight of professional investors and negative coefficient. All other Google searches are considered to have an overweight of household investors.

Google searches	***N***	**Tested time lags**	**Mean**	**Median**	**St. dev.**	**Min**	**Max**
First section, rumor, and news searches							
Apple Rumors	**260 weeks**	**1–4 weeks**	**0.9**	**1.0**	**0.4**	**0.5**	**3.0**
9to5mac	260 weeks	1–4 weeks	1.2	1.0	0.6	0.5	5.0
TheVerge	260 weeks	1–4 weeks	1.0	1.0	0.2	0.5	2.0
MacRumors	**260 weeks**	**1–4 weeks**	**2.7**	**2.0**	**1.5**	**1.0**	**12.0**
AppleInsider	260 weeks	1–4 weeks	0.7	0.5	0.3	0.5	2.0
Second section, Apple, and related big tech stock searches							
AAPL	**260 weeks**	**1–4 weeks**	**20.2**	**18.0**	**7.6**	**9.0**	**51.0**
AMZN	**260 weeks**	**1–4 weeks**	**14.9**	**13.0**	**9.7**	**2.0**	**48.0**
IBM	260 weeks	1–4 weeks	72.2	72.0	10.9	38.0	100.0

Table 2: Descriptive statistics for the Google searches

From Table 2, the search for "IBM" has the highest index number 100, all other Google searches being indexed to it. Had this search not been included, the search for AAPL would have been the main index, meaning all other Google searches

would have been indexed after it, since it has the second highest index of 51. Excluding the Google search on IBM, wouldalso have increased the index numbers for Google searches on MacRumors, Apple rumors, and AMZN, which could have changed their significance.

In Autometrics, the estimation sample cannot start before week 19 in 2015 because of the tested time lags for 1–4 weeks in the dataset. Therefore, the Autometrics estimation is conducted from week 19 in 2015 to week 42 in 201, covering 88% of the dataset. The last 26 weeks of the dataset from week 43 in 2019 to week 16 in 2020 are chosen as hold-out data for the forecast evaluation.

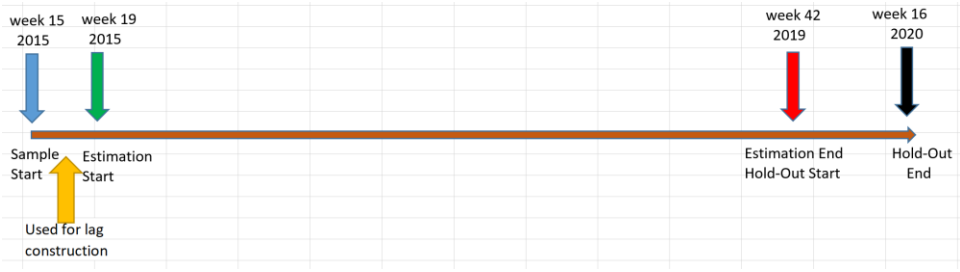


Fig. 6: Timeline for the dataset and visualization of the train/test split.

Fig. 7 shows time plots of the two tested dependent variables, namely the weekly volatility and first difference log(avg close), which is the stock price return. It also shows the time plots of the two most significant regressors, the Google searches for AAPL and AMZN and the highest Google search index in the 5-year period for IBM.

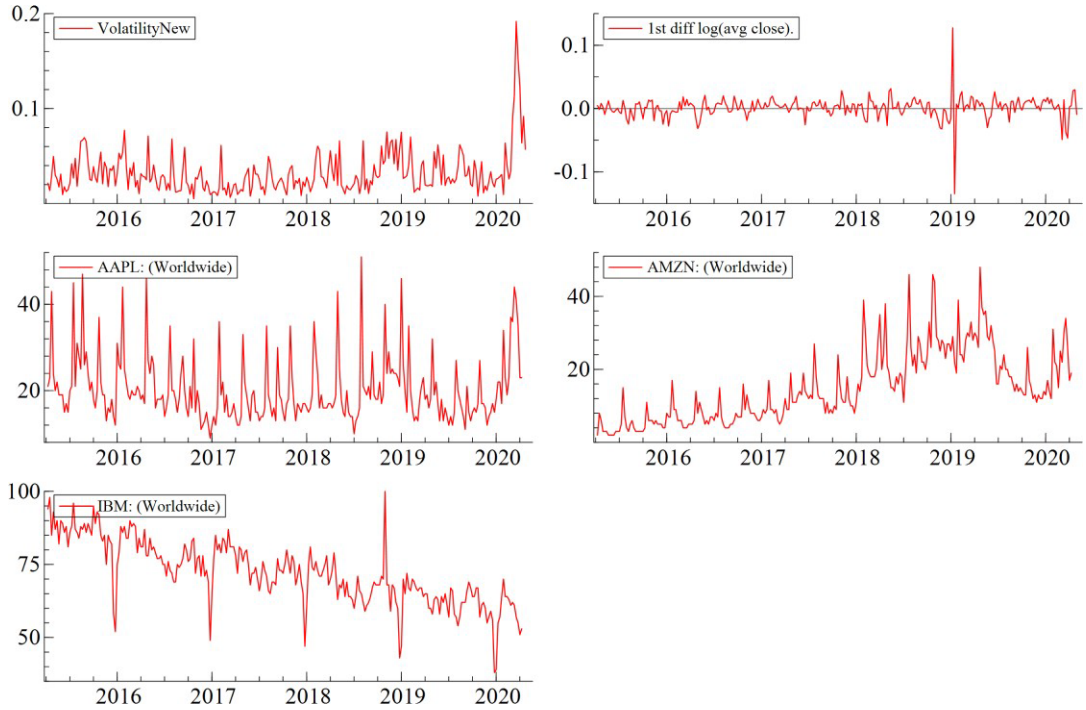


Fig. 7: Timeplots of the main variables

The selected Google searches developed as follows:

AAPL: Quarterly peaks around the ends of January, April, July, and October in each year, when Apple is releasing its quarterly reports to Nasdaq and on their investor site.

AMZN: Quarterly peaks around the ends of January, April, July, and October in each year, when Amazon is releasing its quarterly reports to Nasdaq and on their investor site.

IBM: Quarterly down peaks around the end of December. An explanation could be that IBM is not linked to the Christmas season, and the search for more Christmas-linked stocks overtakes the December searches. The downward pattern in the Google searches follows the decline in stock prices during 2015–2020. The main role of the Google searches for "IBM" is that it is the highest Google search index in the 5-year period for 10 out of 60 Google searches that were most significant for the weekly Apple stock volatility. Therefore, all last 10 Google

searches in the last dataset tested are indexed after the IBM searches. The declining trend for IBM searches during 2015–2020 follows the declining trend in both IBM stock price and IBM's position in machine learning and AI, where IBM is not among the leaders.

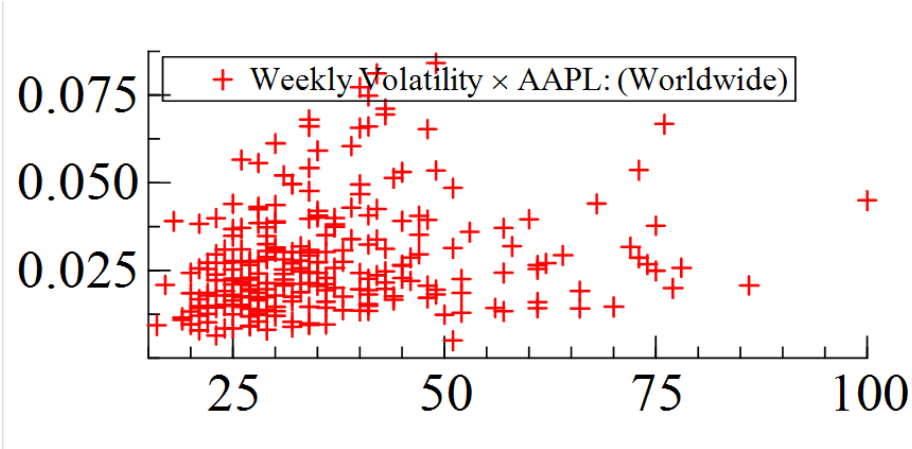


Fig. 8: x-y plot of weekly volatility x AAPL

Y axis is Weekly Volatility and X axis is AAPL Google searches. The x-y plot shows patterns of higher volatility for higher AAPL Google search indexes.

V.5.2. Apple announcement data

Apple releases its quarterly reports at the end of the month after a quarter closes, that is, approximately 1 month after the quarter closes, to stock holders and the media through the investor portal at Apple.com, at https://investor.apple.com/investor-relations/default.aspx.

There was a test of event variables for the iPhone launch in September; the quarterly Reports every end of January, April, July, October; and the Black Friday and Christmas sales. These event variables did not show any effect on either weekly volatility or stock price returns. These event variables were defined as 0/1 variables, taking 1 in the week the event occurred, and 0 otherwise.

There are patterns in some Apple-related Google searches before, during, and after

the quarterly reports, which are considered as quarterly regular spikes in these Google searches. Refer to Fig. 7 for an example of these patterns.

Apart from the quarterly reports from Apple, the single most important event for Apple is the yearly iPhone launch in September. At the iPhone launch 2015, the weekly Apple stock volatility increased, but from 2016 onwards the iPhone launch has been a mean event based on weekly volatility. That also tells a story of the iPhone hype wearing off. The iPhone hype topped in 2012, and has been decreasing since then. Refer to the Google searches for iPhone from 2007–2021:

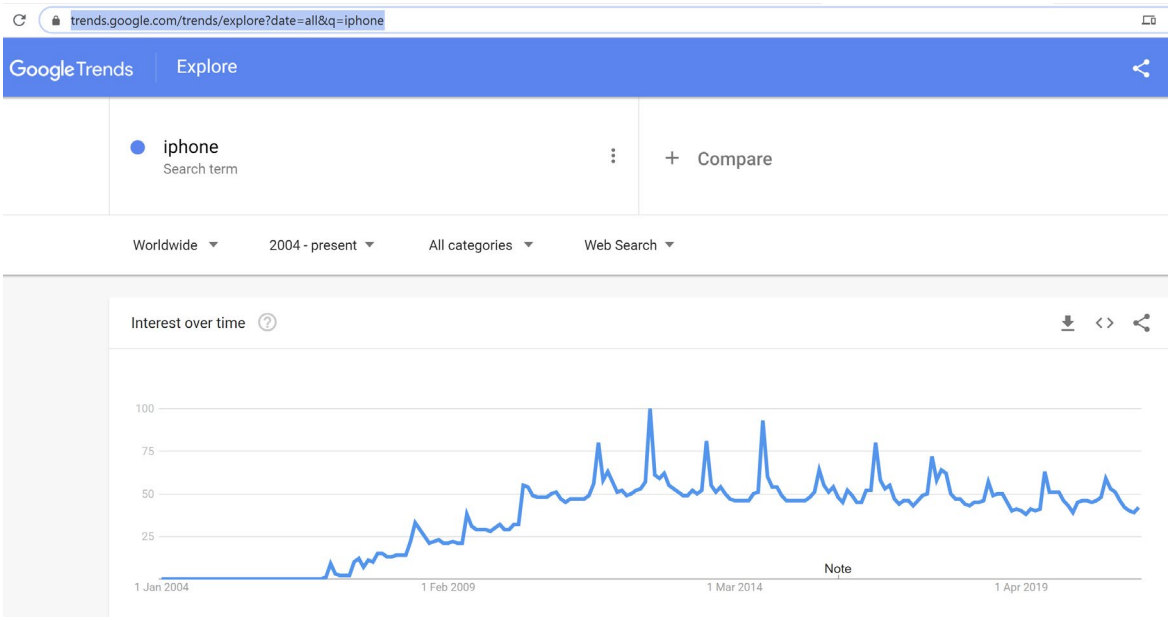


Fig. 9: iPhone Google searches, 2007-2021.

Source: https://trends.google.com/trends/explore?date=all&q=iphone

The first iPhone was launched in June 29, 2007 and, from 2012, the main iPhone launch has always been in September. There are patterns in some Apple-related Google searches before, during, and after the September iPhone launch, which are seen as yearly peaks in these searches. Refer to Fig. 7 for an example of these patterns.

There are also some announcement data on other Apple products, but the most

important product announcement for Apple and Apple stock is theyearly iPhone launch.

V.5.3. Weekly return data

Nasdaq.com provides free data download for their listed stocks going back10 years at https://www.nasdaq.com/symbol/aapl/historical.

These data provide daily info for the selected period on Apple stockprice, namely open, high, low, close, and volume. These daily data from Nasdaq.com were the basis for calculating weekly Apple stock price volatility, based on formula (5).

These daily data were also used to calculate the average weekly stock priceand weekly volatility. The weekly average stock price was calculated based on the average of all the daily closing prices. Daily closing prices were also used to calculate the daily changes, forming the basis for calculating the weekly volatility. As the Nasdaq data only have daily close, high, low values for stock prices—with no daily average—it was chosen to use the close for the weekly calculations of the financial variables. These weekly financial variables were needed in the model together with the weekly Google searches.

V.5.4. Formulas

$\boldsymbol{P_t}$ is the weekly average stock price in week t, calculated as:

(1) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}} + \boldsymbol{P_{4t}} + \boldsymbol{P_{5t}})/5$,

where $\boldsymbol{P_{it}}$ is the closing stock price for Apple on day i in week t.

Approximately 80% of the trading weeks, have 5 trading days—Monday to Friday—based on the above formula.

The remaining 20% of the trading weeks have 4 trading days, the formula being:

(2) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}} + \boldsymbol{P_{4t}})/4$.

Very few trading weeks have 3 trading days, with the following formula:

(3) $\boldsymbol{P_t} = (\boldsymbol{P_{1t}} + \boldsymbol{P_{2t}} + \boldsymbol{P_{3t}})/3$.

The stock price return in this article is expressed as a first difference log variable,in line with the ARCH models. The use of stock price log returns has advantages over the arithmetic return (see, e.g.. Hudson & Gregoriou 2010). The first difference stock price return is expressed as:

(4) $\Delta \boldsymbol{log}(\boldsymbol{P_t}) = \boldsymbol{log}(\boldsymbol{P_t}) - \boldsymbol{log}(\boldsymbol{P_{t-1}})$.

The best explanatory variable for the weekly stock price return is theweekly stock price return from the previous week and two weeks before. Prior to volatility, I also investigated modeling stock price return. Theresults are available upon request. The historical stock price return only creates a R^2 of 10% for the training data and 6% for the hold-out data. With the adding of the best three Google searches, R^2 increases to 17% for the training data and 10% for the hold-out data. The model for stock price return is not strong, the focus of this study being thus on modeling the weekly volatility for the Apple stock.

For the weekly average volatility, the model is much stronger. I follow Christiansen et al. (2012) and Paye (2012), and define weekly volatility as:

(5) $$\boldsymbol{\sigma_t} = \sqrt{(\boldsymbol{r_{1t}}^2 + \boldsymbol{r_{2t}}^2 + \boldsymbol{r_{3t}}^2 + \boldsymbol{r_{4t}}^2 + \boldsymbol{r_{5t}}^2)},$$

where the *r*s are the five daily changes in stock price for a week with 5 trading days. For 80% of the weeks in the dataset with five trading days, there are five *r*s

in the above formula. For 20% of the weeks in the dataset with 4 trading days, there are four *r*s in the formula. For the few weeks with 3 trading days,there are three *r*s in the formula. A classical weekly volatility formula is the standard deviation of five dailystock price changes, which is a variance. The above formula does not subtract the mean from each daily change, before considering each of the five daily returns in a week. The above formula is also not dividing with N - 1 before applying the square root. Therefore, compared to the classical volatility formula, the new formula can be interpreted as a measure of stock price fluctuations in a given week without using the weekly mean for daily changes.

All variables are now defined, so the final regression model for weeklyvolatility can be defined as:

(6) $\sigma_t = \alpha_0 + \alpha_1 \sigma_{t-1} + \cdots . + \alpha_p \sigma_{t-p} + \sum^{N} \beta_i X_{it}^{i=} + \varepsilon_t,$

where X is the explanatory variable constructed from the Google search data and also the event variables for iPhone launch, Black Friday andChristmas sales, and quarterly reports. The event dummies are defined as0/1 variables, which take 1 in the week the event occurred, and 0 otherwise.

V.6. Results and Discussion

The initial test runs on 62 Google searches not indexed against each other, were mostly used to identify the most relevant Google searches. After the most relevant 10 Google searches were indexed against each other, both IBM SPSS LASSO and Autometrics picked only two relevant Google searches, that is, “AAPL” and AMZN,” which are the stock symbols for Apple and Amazon, respectively.

Oxmetrics 8.10 was used to model the 62 Google searches as input variables, with weekly volatility as the dependent variable. Using the automatic model selection function in Oxmetrics 8.10, called Autometrics, the 62 Google searches were tested

with time lags from 1 to 4 weeks, and 51 weekly seasonal dummies were also included in the modelling. Event dummies for Apple quarterly reports, iPhone releases, and Black Friday and Christmas sales were also tested.

The final model output in Table 3 is from the automatic model selection, single-equation dynamic modeling using Autometrics.

Software:	OxMetrics 8.10					IBM SPSS Statistics 26				
Variable selection method:	Autometrics					Lasso				
Variables selected	Coefficient	Std.Error	t-value	t-prob	Part.R^2	Coefficient	Std.Error	t-value	t-prob	Part.R^2
t-1 AAPL: (Worldwide)	0.00147	0.0890	16.5	0.000	0.5552	0.00132	0.0594	22.2	0.000	0.6901
t-4 AMZN: (Worldwide)	0.00037	0.0763	4.91	0.000	0.0995	0.00027	0.0715	3.71	0.000	0.0585
t-4 macrumors: (Worldwide)	0.00183	0.0008	2.39	0.018	0.0255					
t-4 apple rumors: (Worldwide)	-0.00659	0.0027	-2.45	0.015	0.0268					
Diagnostics tests:						**Diagnostics tests:**				
AR 1-7 test:	F(7,211) = 2.6487 [0.0121]*					AR 1-7 test: F(7,215) = 4.2173 [0.0002]**				
ARCH 1-7 test:	F(7,210) = 4.4896 [0.0001]**					ARCH 1-7 test: F(7,210) = 3.0202 [0.0048]**				
Normality test:	Chi^2(2) = 9.5215 [0.0086]**					Normality test: Chi^2(2) = 10.519 [0.0052]**				
Hetero test:	F(10,213) = 3.4856 [0.0003]**					Hetero test: F(4,219) = 7.9119 [0.0000]**				
Hetero-X test:	F(25,198) = 3.1901 [0.0000]**					Hetero-X test: F(5,218) = 6.7054 [0.0000]**				
RESET23 test:	F(2,216) = 6.7985 [0.0014]**					RESET23 test: F(2,220) = 5.8947 [0.0032]**				

Table 3: Model Output from Autometrics & Lasso

V.6.1. Diagnostic tests

The AR 1-7 test is a standard test of autocorrelation up to degree 7. It teststhe joint hypothesis that εˆt is uncorrelated with εˆt−j, for any choice of j, against the alternative that εˆt is correlated with εˆt−j. The null hypothesis of no autocorrelation between residuals can be rejected for a P-value of 1.21% and significance level of 2.5%. Therefore, there is formal evidence of little autocorrelation between residuals. At a significance level of 1%, the null hypothesis is accepted, which makes it a borderline scenario for this test. In the LASSO model with just two predictors, there is clear rejection of thenull hypothesis and formal evidence of autocorrelation between residuals.

The ARCH 1-7 test is a standard ARCH test of the null hypothesis of no ARCH effect, that is, if the squared standardized residuals do not exhibit autocorrelation. The null hypothesis of no ARCH effect is rejected for P-values of 0.0001 and 0.0048. Therefore, there is formal evidence of the ARCH effect in the model.

Normality test. The null hypothesis of normality is rejected at a significance level of 1%, with P-values of 0.86% and 0.52%. Hence, there is no formal evidence of normality for this model.

Hetero and hetero-X tests. The null hypothesis of homoscedasticity can be rejected for P-values of 0–0.03% at the 1% significance level. Hence, there is formal evidence of heteroscedasticity in this model.

RESET123 test. The regression specification error test has a null hypothesis of no squared and cubic terms in the regression model. The nullhypothesis can be rejected for P-values of 0.14% and 0.32% at the 1% significance level. Hence, there is formal evidence of mis-specification of the regression model from this test.

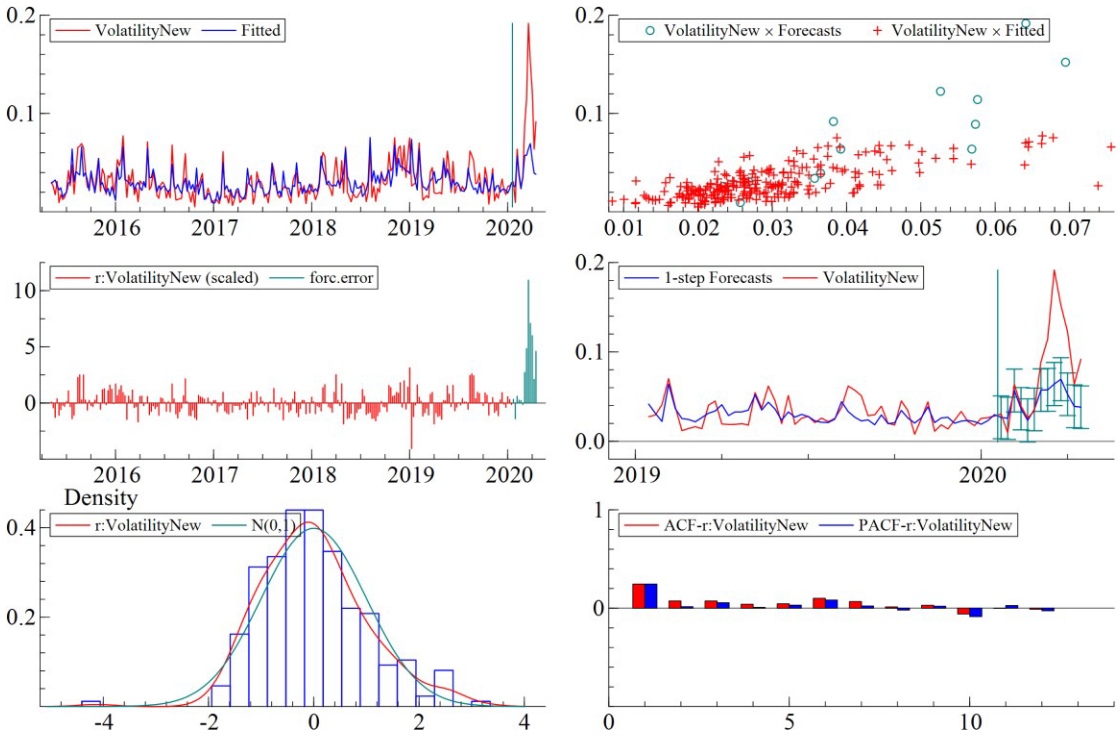


Fig. 10: Model Output from Autometrics

The residual plot shows a random pattern, suggesting a linear model wouldfit well the data. Residuals are also normally distributed, again suggesting the linear regression model is fitting the dependent variable well. The ACF and PACFplots show no autocorrelation. The COVID-19 peak in March 2020 is an outlier, but the model is not capturing this, as there were no COVID-19 data for training the model.

V.7. Forecasting evaluation

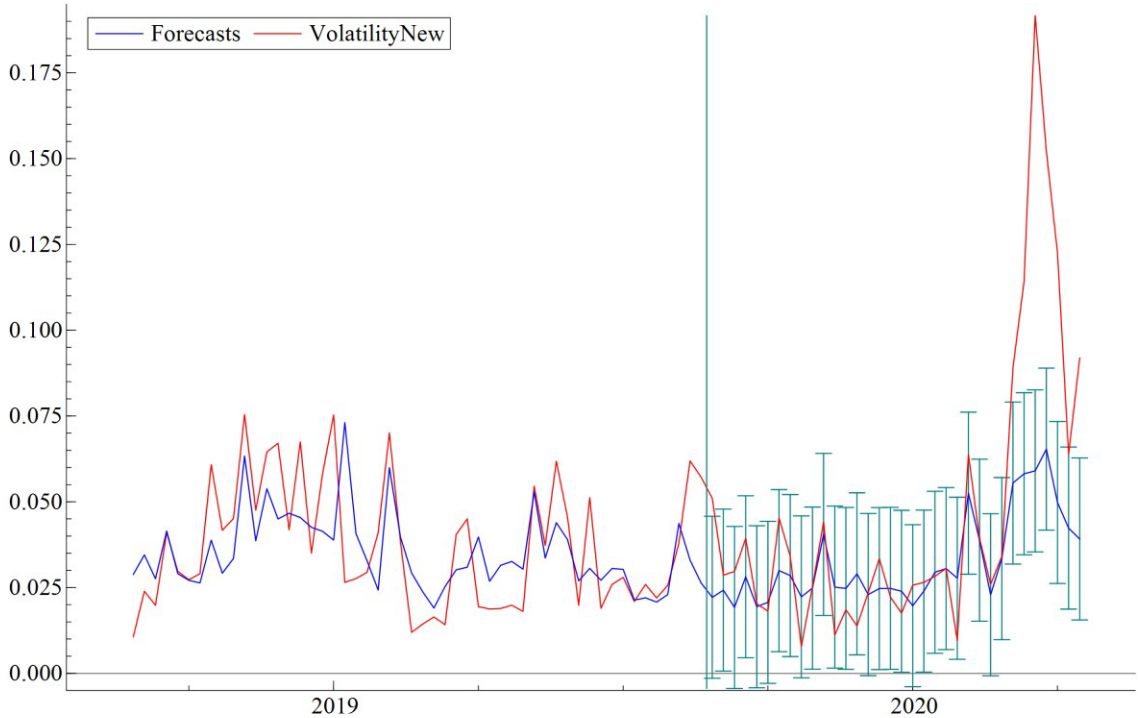


Fig. 11: Forecast graph, Y axis is Weekly volatility. 34 weeks out-of-sample

In Fig. 11, the blue line is the out-of-sample forecast from September 2019 to April 2020 for the last 34 weeks of the dataset. The estimation sample is from April 2015 to August 2019. The green band around the blue line is a 95% band marking +/- 2 forecast standard errors.

From late February 2020 to April 2020, the weekly volatility for the Apple stock

took a hit due to the COVID-19 pandemic period, and the forecast model fail to forecast this peak. This is probably due to the COVID-19 pandemic not being captured by the Google searches in this model. The Oxmetrics output for the above 34 weeks dynamic forecast out-of-sample is in Appendix 1, including the COVID-19 pandemic period.

To test how the forecasting performed when excluding the COVID-19 pandemic period, a forecast scenario similar to the setting was tested for a 26-week forecast ending forecast in February 2020 before the pandemic hit the stock market. The difference in forecasting periods for the 34 weeks including the COVID-19 pandemic period and 26 weeks excluding it ensures identical training datasets from week 19 in 2015 to week 34 in 2019 under both forecasting scenarios.

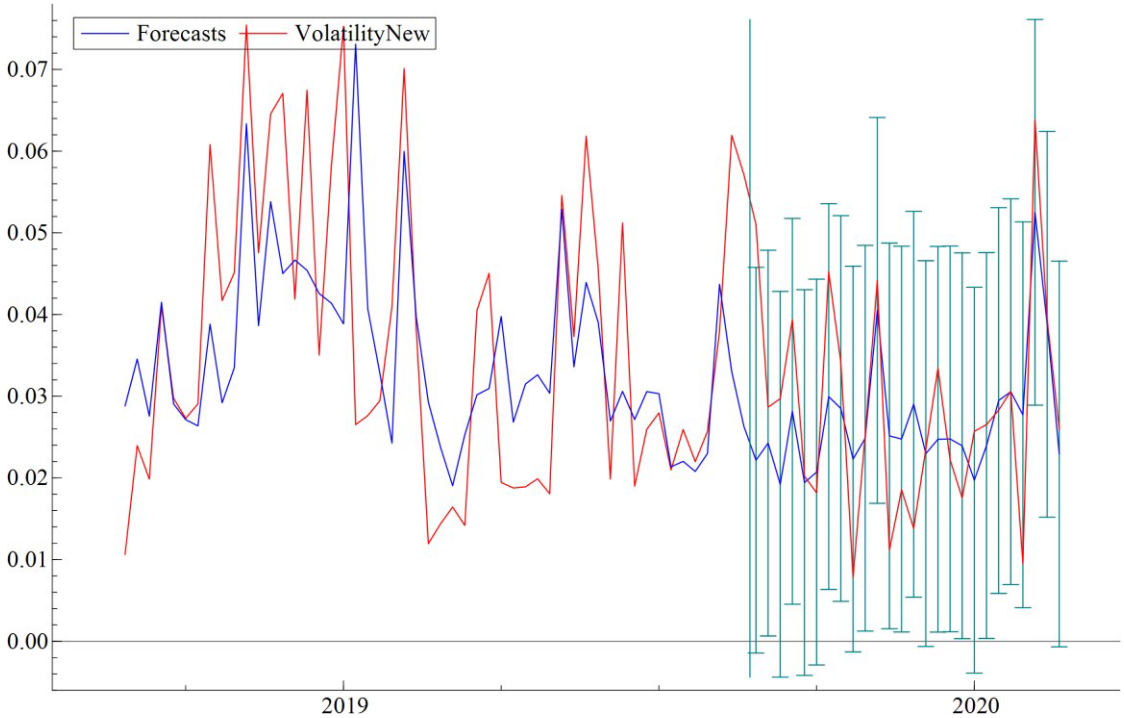


Fig. 12: Forecast graph, Y axis is Weekly volatility. 26 weeks out-of-sample

The Oxmetrics output for the above 26 weeks out-of-sample dynamic forecast is shown in Appendix 2. The Oxmetrics outputs for the two out-of-sample dynamic forecasts without and with the COVID-19 pandemic period are summarized in

Table 4.

Excluding the COVID-19 pandemic period		Including the COVID-19 pandemic period	
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35) –2020 (16)	
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)	
Parameter constancy forecast tests:		Parameter constancy forecast tests:	
Forecast Chi^2(26) = 19.244 [0.8259]		Forecast Chi^2(34) = 291.67 [0.0000] **	
Chow F(26,218) = 0.73626 [0.8217]		Chow F(34,218) = 7.5341 [0.0000] **	
CUSUM t(25) = 0.4978 [0.6230]		CUSUM t(33) = 6.580 [0.0000]**	
RMSE =	0.010157	RMSE =	0.034577
MAPE =	38.524	MAPE =	40.205
mean(Error)=	0.001264	mean(Error)=	0.014414
SD(Error)=	0.010078	SD(Error)=	0.031434

Table 4: Autometrics Output excluding & including COVID-19.

Comparing the forecasts without and with the pandemic period, the former performs relatively well in terms of Chi^2, Chow, CUSUM, RMSE, MAPE, and mean(Error).

All forecasting was conducted using the model with four predictors selected by Autometrics. For comparing the two used variable selection methods in this article—Autometrics and LASSO—Table 5 presents both methods and their forecasting KPIs.

Software:	OxMetrics 8.10	IBM SPSS Statistics 26
Variable selection method:	**Autometrics**	**Lasso**
Variables selected: t-1 AAPL, t-4 AMZN, t-4 macrumors & t-4 apple rumors		t-1 AAPL & t-4 AMZN
Forecasting excluding the COVID-19 pandemic period		**Forecasting excluding the COVID-19 pandemic period**
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35)–2020 (8)
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)
Parameter constancy forecast tests:		Parameter constancy forecast tests:
Forecast Chi^2(26) = 18.646 [0.8510]		Forecast Chi^2(26) = 19.189 [0.8284]
Chow F(26,218) = 0.71147 [0.8482]		Chow F(26,222) = 0.73701 [0.8210]
CUSUM t(25) = 0.6245 [0.5380] (zero forecast innovation mean)		CUSUM t(25) = 0.3433 [0.7342]
Forecasting including the COVID-19 pandemic period		**Forecasting including the COVID-19 pandemic period**
One-step (ex post) forecast analysis 2019 (35)–2020 (8)		One-step (ex post) forecast analysis 2019 (35)–2020 (8)
Training dataset 2015 (19)–2019 (34)		Training dataset 2015 (19)–2019 (34)
Parameter constancy forecast tests:		Parameter constancy forecast tests:
Forecast Chi^2(34) = 288.54 [0.0000]**		Forecast Chi^2(34) = 278.85 [0.0000]**
Chow F(34,218) = 7.4183 [0.0000]**		Chow F(34,222) = 7.6773 [0.0000]**
CUSUM t(33) = 6.659 [0.0000]** (zero forecast innovation mean)		CUSUM t(33) = 6.536 [0.0000]**

Table 5: Forecasting output from Autometrics & SPSS

V.7.1. Autometrics recursive graphs

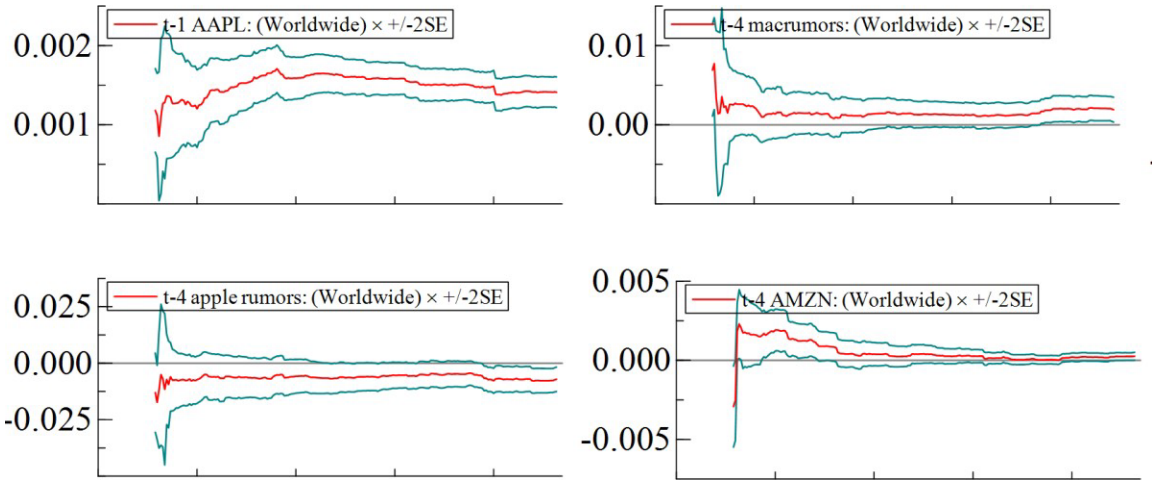


Fig. 13: Graphs of the coefficients for the most important predictors

In Fig. 13, the Google searches for AAPL have the most significant pattern. The coefficient on the Google search AMZN t - 4 is close enough to zero so that it is not worth analyzing it. However, the IBM SPSS LASSO has a much higher partial R^2 on 8% for this variable compared to only 2% in Autometrics. The coefficients for the Google searches for MacRumors t - 4 and Apple rumors t - 4 are larger than for AAPL, but these Google searches are also relative small compared to the AAPL searches.

V.7.2. IBM SPSS 26 output

Model Summary

Multiple R	R Square	Adjusted R Square	Regularization "R Square" (1-Error)	Apparent Prediction Error	Expected Prediction Error		
					Estimate[a]	Std. Error	N[b]
,725	,525	,514	,475	,525	,560	,077	256

Penalty ,420

Dependent Variable: VolatilityNew

Predictors: t-1 AAPL: (Worldwide) t-1 9to5mac: (Worldwide) t-1 theVerge: (Worldwide) t-1 macrumors: (Worldwide) t-1 IBM: (Worldwide) t-1 appleinsider: (Worldwide) t-1 AMZN: (Worldwide) t-1 apple rumors: (Worldwide) t-1 SSNLF: (Worldwide) t-2 AAPL: (Worldwide) t-2 9to5mac: (Worldwide) t-2 theVerge: (Worldwide) t-2 macrumors: (Worldwide) t-2 IBM: (Worldwide) t-2 appleinsider: (Worldwide) t-2 AMZN: (Worldwide) t-2 apple rumors: (Worldwide) t-2 SSNLF: (Worldwide) t-3 AAPL: (Worldwide) t-3 9to5mac: (Worldwide) t-3 theVerge: (Worldwide) t-3 macrumors: (Worldwide) t-3 IBM: (Worldwide) t-3 appleinsider: (Worldwide) t-3 AMZN: (Worldwide) t-3 apple rumors: (Worldwide) t-3 SSNLF: (Worldwide) t-4 AAPL: (Worldwide) t-4 9to5mac: (Worldwide) t-4 theVerge: (Worldwide) t-4 macrumors: (Worldwide) t-4 IBM: (Worldwide) t-4 appleinsider: (Worldwide) t-4 AMZN: (Worldwide) t-4 apple rumors: (Worldwide) t-4 SSNLF: (Worldwide)

a. .632 Bootstrap estimate (50 bootstrap samples).

b. If N is smaller than the number of active (training) cases, this is due to excluding cases from estimation of the expected prediction error for reason(s) explained in the warnings table.

Table 6: Model summary from SPSS Lasso

The LASSO model in IBM SPSS 26 has an R^2 on 52.5%, which is very comparable to the one in Autometrics, around 50%. The above lambdian model yielded 0.42, which is optimized after the 19% bootstrap test set (50 bootstrap samples out of the 260 weekly observations).

V.8. Final model

The automatic model selection in Autometrics results in two significant Google searches linked to the weekly volatility of the Apple stock and no effects from the seasonal and event dummies or historical values of volatility. The significant Google searches are AAPL t - 1 and AMZN t - 4.

The Google search **"AAPL"** is time lagged 1 week, being the most significant search linked Apple stock volatility, with a partial R^2 of 49% in Autometrics and 50% in SPSS 26 LASSO.

The Google search AAPL at t - 1, which is the Google search for the Apple

stock symbol on Nasdaq, is driving up volatility 1 week after thesearches. This is because it is assumed the Google searches for AAPL are mainly done by private investors, under the assumption that professional investors have a lower need to Google search the Apple stock symbol AAPL. These assumptions cannot be proven but are logical.

To verify this conjecture, I interviewed Henrik Ekman, Independent Investment Consultant, former Head of Equities at Maj Invest, on January 28, 2021. He confirmed that professional portfolio analysts are using Google searches to find indications of sales going up and down for the stocks they are analyzing. They also use Google searches in general for information gathering for the stocks currently in their portfolio. While I did not find out any specifics on the difference between private and professional investors in terms of Google searches for stock symbols, this confirmed the professional investors' general use of Google searches, as shown in this articles Investor Journey Model.

The LASSO algorithm was run in IBM SPSS 26 on the exact same dataset, with weekly volatility as the dependent variable and two significant input variables. **AAPL was** time lagged 1 week as the most significant input variable, similar to the Autometrics method, with a similar partial R^2 of 50%. For the second significant input variable, LASSO also chose **AMZN** t – 4, with a partial R^2 of 8%. In Autometrics, AMZN t - 4 had a similar partial R^2.

Autometrics was also tested with a target size of 5% instead of the 1% target size for all other tests. The target size is the t-probability threshold for choosing and eliminating input variables.

The target size for 5% also results in the choice of the MacRumors t - 4 Google search with a partial R^2 of 2.7% and a positive coefficient and Apple rumors t - 4 with a partial R^2 of 3.1% and negative coefficient. Given these relative small partial

R^2, further analysis is not necessary. The positive coefficient on MacRumors indicates more private investors conduct these Google searches, while the negative coefficient on Apple rumors indicates more professional investors rely on these Google searches. MacRumors.com has good reputation for vetting Apple rumors, but the Google searches for Apple rumors lead also to MacRumors.com, 9to5mac.com, AppleInsider.com, and TheVerge.com; this wider mix of Apple rumor news sites are used more by professional investors compared to just MacRumors.com. However, given the small partial R^2 of around 3% for these Apple rumor Google searches, this analysis should of course be interpreted cautiously.

V.9. Conclusions

More than 60 Apple-related Google searches were tested in this study aspredictors for weekly Apple stock volatility. Under the framework of the newly proposed investor journey model, I analyzed and explained why a buzz in some Apple-related Google searches will dampen the weekly Apple stock volatility and why a buzz in the other searches will increase the weekly volatility for the Apple stock.

A buzz up in the Google search for "AAPL" will increase the Apple stock price volatility in the following week. This is the most significant pattern, since the partial R^2 for the Google search "AAPL" is 44% in Autometrics and 50% in SPSS's LASSO.

A buzz up in Google search "AMZN" will also increase Apple stock price volatility after 4 weeks. However, the effect is small compared to AAPL, as the partial R^2 for this Google search is just 2% in Autometrics and 8% in SPSS LASSO. With the small partial R^2 of 2–8% for these AMZN Google searches, the results of this analysis should be interpreted with caution.

When the target size changed from 1% to 5% in Autometrics, which is the t-probability threshold for choosing and eliminating input variables, MacRumors t - 4 and Apple rumors t - 4 also become significant, but their partial R^2 are 2.7% and 3.1%, respectively, so their effect is quite small. MacRumors has a positive coefficient and will increase the Apple stock price volatility after 4 weeks. Apple rumors has a negative coefficient and will decrease the Apple stock price volatility after 4 weeks.The explanation could be that an increase in Google searches 4 weeks before this information is available will result in investors buying and selling more. However, it could also be a random pattern, considering the small partial R^2 of around 3% for these Apple rumor news Google searches.

The most predictive Google search for Apple stock volatility was AAPL t-1. The Google searches for AAPL stock symbol are mostly done by private investors, which explains why their buzz up increases volatility.

When the COVID-19 pandemic disrupted the stock market during February–April 2020, both private and professional investors panicked, which is why and the proposed model could not capture the disruption, as the model was not trained with data from this period.

Further research ideas would be to model the Amazon stock based on Google searches, as the AMZN Google searches show a better visual correlation with the Amazon stock price compared to the Apple stock and AAPL Google searches. Perhaps the stock price return for the Amazon stock has a better potential for being modeled and predicted with Google searches. However, there is the need to test whether the Amazon stock price volatility modeling based on Google searches is stronger than that for Apple in this article. That could require an additional model.

The novel investor journey model presented in this article will thus enable further analyses linking big social data to investor behavior on the financial markets.

V. References

Asur, S., Huberman, B. A., 2010. Predicting the future with social media. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, pp. 492–499.

Boldt, L. C., Vinayagamoorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A., Vatrapu, R., 2016.Forecasting Nike's sales using Facebook data. 2016 IEEE International Conference on Big Data (Big Data), pp. 2447–2456.

Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of Computer Science, 2*(1), pp. 1–8.

BusinessInsider.com, April 2018. How Google retains more than 90% of market share.. http://uk.businessinsider.com/how-google-retains-more-than-90-of-market-share-2018-4?r=US&IR=T

Buus Lassen, N., la Cour, L., Vatrapu, R., 2017. Predictive analytics with social media data. In Sloan, L. and Quan-Haase, A. (eds). *The SAGE Handbook of Social MediaResearch Methods*, Chapter 20, pp. 328–341, Sage.

Christiansen, C., Schmeling, M., Schrimpf, A., 2012. A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, *27*(6), pp. 956–977.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), pp. 253–263. doi:10.2307/1392185

Federal Reserve Board, 2016. Triennial Survey of Consumer Finance (SCF). https://www.federalreserve.gov/econres/files/BulletinCharts.pdf

Greenwich Associates, 2015. Social media influencing investment decisions at globalinstitutions. https://www.greenwich.com/press-release/social-media-influencing-investment- decisions-global-institutions

Hendry, D., Castle, J., Doornik, J., Johansen, S., Pretis, F., 2014. Model selection with big data. Oxmetrics Conference, September 2014. http://www.timberlake.co.uk/media/wysiwyg/pdf/David%20F.%20Hendry.pdf

Hendry, D. F., Doornik, J.A., 2014. Empirical model discovery and theory evaluation: Automatic model selection methods in Econometrics.

Hudson, R., Gregoriou, A., 2010. Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. Available at SSRN: https://ssrn.com/abstract=1549328 or http://dx.doi.org/10.2139/ssrn.1549328

Jacobsen, P. Ø., Ringberg, T., Bjerre M., CBS & CBS BIG DATA LAB RF2016-2017 + interviews 2018 inspired from: Customer Infinity Model – Per Østergaard Jacobsen, Mogens Bjerre, Claus Andersen 2015. https://cbs-executive.dk/wp-content/uploads/2019/05/Ten-Deadly-Marketing-Sins-30-april-2019-report-and-conclusions.pdf

Jiao, P., Veiga, A., Walther, A., September 25, 2018. Social media, news media and the stock market. Available at: https://ssrn.com/abstract=2755933 or http://dx.doi.org/10.2139/ssrn.2755933

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. *Econometrica, 47*, pp. 263–291.

Lassen, N. B., Madsen, R. and Vatrapu, R., 2014,. Predicting iPhone sales from

iPhone tweets. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, IEEE, pp. 81–90.

Levy, M., Levy, H., 2002. Prospect theory: Much ado about nothing?. *Management Science, INFORMS*, *48*(10), pp. 1334–1349. DOI: 10.1287/mnsc.48.10.1334.276

Levy, H., 2015. *Stochastic Dominance: Investment Decision Making under Uncertainty.* Springer.

Li, T., van Dalen, J., van Rees, P. J., 2018. More than just noise? Examining the information content of stock microblogs on financial markets. *Journal of InformationTechnology*, *33*(1), pp. 50–69.

MarketWatch.com, 2018. Finance Twitter: The 50 most important people for investors to follow. https://www.marketwatch.com/story/finance-twitter-the-50-most-important-people-for- investors-to-follow-2018-12-13

Nwogugu, M., 2005. Towards multi-factor models of decision making and risk: A critique of Prospect Theory and related approaches, part I. *Journal of Risk Finance*, *6*(2), pp. 150–162. https://doi.org/10.1108/15265940510585815

Paye, B.S., 2012. "Déjà vol": Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, *106*(3),pp. 527–546.

Preis, T., Moat, H. S., Stanley, H. E., 2013. Quantifying trading behavior in financial markets using Google Trends. *Nature, Scientific Reports*, *3*, p. 1684. DOI:10.1038/srep01684 (2013).

TheGlobeAndMail.com, 2018. How Twitter can help institutional investors

market better trading. https://www.theglobeandmail.com/business/careers/business-education/article-how- twitter-can-help-institutional-investors-make-better-trading/

Wakker, P. P., 2010. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press.

V. Appendix 1. Forecast output from Oxmetrics, September 2019–April 2020, including the COVID-19 pandemic period

```
1-step (ex post) forecast analysis 2019(35) - 2020(16)
Parameter constancy forecast tests:
Forecast  Chi^2(34) =   291.67 [0.0000]**
Chow      F(34,218) =   7.5341 [0.0000]**
CUSUM     t(33)     =    6.580 [0.0000]**  (zero forecast innovation mean)

Dynamic (ex ante) forecasts for VolatilityNew (SE based on error variance only)
   Horizon      Forecast        SE        Actual       Error   t-value        -2SE        +2SE
  2019(35)     0.0221757   0.01181     0.0510110    0.028835     2.443  -0.0014354    0.045787
  2019(36)     0.0242622   0.01181     0.0286650   0.0044028     0.373  0.00065115    0.047873
  2019(37)     0.0192232   0.01181     0.0297024    0.010479     0.888  -0.0043879    0.042834
  2019(38)     0.0281562   0.01181     0.0393877    0.011231     0.951   0.0045452    0.051767
  2019(39)     0.0194324   0.01181     0.0202125  0.00078013     0.066  -0.0041787    0.043043
  2019(40)     0.0207155   0.01181     0.0181770  -0.0025385    -0.215  -0.0028956    0.044327
  2019(41)     0.0299563   0.01181     0.0452540    0.015298     1.296   0.0063453    0.053567
  2019(42)     0.0284997   0.01181     0.0341143   0.0056147     0.476   0.0048886    0.052111
  2019(43)     0.0223038   0.01181    0.00787080   -0.014433    -1.223  -0.0013072    0.045915
  2019(44)     0.0248636   0.01181     0.0253074  0.00044384     0.038   0.0012525    0.048475
  2019(45)     0.0405084   0.01181     0.0441805   0.0036721     0.311    0.016897    0.064119
  2019(46)     0.0251504   0.01181     0.0111957   -0.013955    -1.182   0.0015393    0.048761
  2019(47)     0.0247585   0.01181     0.0185567  -0.0062018    -0.525   0.0011474    0.048370
  2019(48)     0.0290163   0.01181     0.0138208   -0.015195    -1.287   0.0054052    0.052627
  2019(49)     0.0229785   0.01181     0.0235305  0.00055208     0.047 -0.00063259    0.046590
  2019(50)     0.0247348   0.01181     0.0334344   0.0086996     0.737   0.0011237    0.048346
  2019(51)     0.0247689   0.01181     0.0222302  -0.0025387    -0.215   0.0011579    0.048380
  2019(52)     0.0239338   0.01181     0.0175467  -0.0063871    -0.541  0.00032276    0.047545
   2020(1)     0.0197210   0.01181     0.0257094   0.0059884     0.507  -0.0038901    0.043332
   2020(2)     0.0239680   0.01181     0.0265275   0.0025596     0.217  0.00035692    0.047579
   2020(3)     0.0294778   0.01181     0.0282960  -0.0011817    -0.100   0.0058667    0.053089
   2020(4)     0.0305566   0.01181     0.0306042  4.7624e-05     0.004   0.0069455    0.054168
   2020(5)     0.0277304   0.01181    0.00949557   -0.018235    -1.545   0.0041193    0.051341
   2020(6)     0.0525129   0.01181     0.0638056    0.011293     0.957    0.028902    0.076124
   2020(7)     0.0388066   0.01181     0.0394267  0.00062008     0.053    0.015196    0.062418
   2020(8)     0.0229329   0.01181     0.0259554   0.0030225     0.256 -0.00067814    0.046544
   2020(9)     0.0334589   0.01181     0.0340971  0.00063827     0.054   0.0098478    0.057070
  2020(10)     0.0554885   0.01181     0.0890427    0.033554     2.842    0.031877    0.079100
  2020(11)     0.0581828   0.01181      0.114265    0.056082     4.750    0.034572    0.081794
  2020(12)     0.0590218   0.01181      0.191717     0.13270    11.240    0.035411    0.082633
  2020(13)     0.0653553   0.01181      0.152223    0.086868     7.358    0.041744    0.088966
  2020(14)     0.0498154   0.01181      0.122598    0.072782     6.165    0.026204    0.073426
  2020(15)     0.0423422   0.01181     0.0638132    0.021471     1.819    0.018731    0.065953
  2020(16)     0.0391658   0.01181     0.0919330    0.052767     4.470    0.015555    0.062777
   mean(Error) =     0.014404    RMSE =     0.034577
   SD(Error)   =     0.031434    MAPE =      40.205
```

V. Appendix 2. Forecast output from Oxmetrics, September 2019–February 2020, excluding the COVID-19 pandemic period

```
1-step (ex post) forecast analysis 2019(35) - 2020(8)
Parameter constancy forecast tests:
Forecast  Chi^2(26) =   19.244 [0.8259]
Chow      F(26,218) =  0.73626 [0.8217]
CUSUM     t(25)     =   0.4978 [0.6230]    (zero forecast innovation mean)

Dynamic (ex ante) forecasts for VolatilityNew (SE based on error variance only)
   Horizon      Forecast        SE        Actual       Error   t-value        -2SE        +2SE
  2019(35)     0.0221757   0.01181     0.0510110    0.028835     2.443  -0.0014354    0.045787
  2019(36)     0.0242622   0.01181     0.0286650   0.0044028     0.373  0.00065115    0.047873
  2019(37)     0.0192232   0.01181     0.0297024    0.010479     0.888  -0.0043879    0.042834
  2019(38)     0.0281562   0.01181     0.0393877    0.011231     0.951   0.0045452    0.051767
  2019(39)     0.0194324   0.01181     0.0202125  0.00078013     0.066  -0.0041787    0.043043
  2019(40)     0.0207155   0.01181     0.0181770  -0.0025385    -0.215  -0.0028956    0.044327
  2019(41)     0.0299563   0.01181     0.0452540    0.015298     1.296   0.0063453    0.053567
  2019(42)     0.0284997   0.01181     0.0341143   0.0056147     0.476   0.0048886    0.052111
  2019(43)     0.0223038   0.01181    0.00787080   -0.014433    -1.223  -0.0013072    0.045915
  2019(44)     0.0248636   0.01181     0.0253074  0.00044384     0.038   0.0012525    0.048475
  2019(45)     0.0405084   0.01181     0.0441805   0.0036721     0.311    0.016897    0.064119
  2019(46)     0.0251504   0.01181     0.0111957   -0.013955    -1.182   0.0015393    0.048761
  2019(47)     0.0247585   0.01181     0.0185567  -0.0062018    -0.525   0.0011474    0.048370
  2019(48)     0.0290163   0.01181     0.0138208   -0.015195    -1.287   0.0054052    0.052627
  2019(49)     0.0229785   0.01181     0.0235305  0.00055208     0.047 -0.00063259    0.046590
  2019(50)     0.0247348   0.01181     0.0334344   0.0086996     0.737   0.0011237    0.048346
  2019(51)     0.0247689   0.01181     0.0222302  -0.0025387    -0.215   0.0011579    0.048380
  2019(52)     0.0239338   0.01181     0.0175467  -0.0063871    -0.541  0.00032276    0.047545
   2020(1)     0.0197210   0.01181     0.0257094   0.0059884     0.507  -0.0038901    0.043332
   2020(2)     0.0239680   0.01181     0.0265275   0.0025596     0.217  0.00035692    0.047579
   2020(3)     0.0294778   0.01181     0.0282960  -0.0011817    -0.100   0.0058667    0.053089
   2020(4)     0.0305566   0.01181     0.0306042  4.7624e-05     0.004   0.0069455    0.054168
   2020(5)     0.0277304   0.01181    0.00949557   -0.018235    -1.545   0.0041193    0.051341
   2020(6)     0.0525129   0.01181     0.0638056    0.011293     0.957    0.028902    0.076124
   2020(7)     0.0388066   0.01181     0.0394267  0.00062008     0.053    0.015196    0.062418
   2020(8)     0.0229329   0.01181     0.0259554   0.0030225     0.256 -0.00067814    0.046544
   mean(Error) =    0.0012644    RMSE =    0.010157
   SD(Error)   =     0.010078    MAPE =      38.524
```

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or, How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning – A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus – Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter. En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος*-tænkning hos Martin Heidegger*

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin – en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management - A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være? Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sense-making in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a Partly-Digital World
Comparing Digital Piracy and Legal Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in organizational decision-making

11. Mia Kaspersen
The construction of social and environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte til velgørenhed og understøtte relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig kontekst, strategisk tilgang, samt anvendte redskaber og teknologier for udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural Production
The case of travel guidebook production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie i det 19. århundrede
Omtydningen af melankolien og manien som bipolære stemningslidelser i dansk sammenhæng under hensyn til dannelsen af det moderne følelseslivs relative autonomi.
En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om selvledelsens paradoksale dynamik og eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis
– barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life
- Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelses-studie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity: Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering: Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig – er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm – A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions – on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A – Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats? A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management
A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management
Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management – Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me
- The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration
A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce: Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'. A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity – How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition. Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsifi cation and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics: -An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
   ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
   Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
   Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
   Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
   Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
   Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
   Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
   The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
   Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
    Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
    Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
    Essays on Crude Oil Tanker Markets

13. Cancan Wang
    Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
    Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
    Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
    Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
    The Comparative Dynamics of Private Governance
    The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
    STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
    Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
    Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

23. Henriette Sophia Groskopff Tvede Schleimann
Creating innovation through collaboration – Partnering in the maritime sector

24. Kristian Steensen Nielsen
The Role of Self-Regulation in Environmental Behavior Change

25. Lydia L. Jørgensen
Moving Organizational Atmospheres

26. Theodor Lucian Vladasel
Embracing Heterogeneity: Essays in Entrepreneurship and Human Capital

27. Seidi Suurmets
Contextual Effects in Consumer Research: An Investigation of Consumer Information Processing and Behavior via the Applicati on of Eye-tracking Methodology

28. Marie Sundby Palle Nickelsen
Reformer mellem integritet og innovation: Reform af reformens form i den danske centraladministration fra 1920 til 2019

29. Vibeke Kristine Scheller
The temporal organizing of same-day discharge: A tempography of a Cardiac Day Unit

30. Qian Sun
Adopting Artificial Intelligence in Healthcare in the Digital Age: Perceived Challenges, Frame Incongruence, and Social Power

31. Dorthe Thorning Mejlhede
Artful change agency and organizing for innovation – the case of a Nordic fintech cooperative

32. Benjamin Christoffersen
Corporate Default Models: Empirical Evidence and Methodical Contributions

33. Filipe Antonio Bonito Vieira
Essays on Pensions and Fiscal Sustainability

34. Morten Nicklas Bigler Jensen
Earnings Management in Private Firms: An Empirical Analysis of Determinants and Consequences of Earnings Management in Private Firms

2020

1. Christian Hendriksen
Inside the Blue Box: Explaining industry influence in the International Maritime Organization

2. Vasileios Kosmas
Environmental and social issues in global supply chains: Emission reduction in the maritime transport industry and maritime search and rescue operational response to migration

3. Thorben Peter Simonsen
The spatial organization of psychiatric practice: A situated inquiry into 'healing architecture'

4. Signe Bruskin
The infinite storm: An ethnographic study of organizational change in a bank

5. Rasmus Corlin Christensen
Politics and Professionals: Transnational Struggles to Change International Taxation

6. Robert Lorenz Törmer
The Architectural Enablement of a Digital Platform Strategy

7. Anna Kirkebæk Johansson Gosovic
Ethics as Practice: An ethnographic study of business ethics in a multinational biopharmaceutical company

8. Frank Meier
Making up leaders in leadership development

9. Kai Basner
Servitization at work: On proliferation and containment

10. Anestis Keremis
Anti-corruption in action: How is anti-corruption practiced in multinational companies?

11. Marie Larsen Ryberg
Governing Interdisciolinarity: Stakes and translations of interdisciplinarity in Danish high school education.

12. Jannick Friis Christensen
Queering organisation(s): Norm-critical orientations to organising and researching diversity

13. Thorsteinn Sigurdur Sveinsson
Essays on Macroeconomic Implications of Demographic Change

14. Catherine Casler
Reconstruction in strategy and organization: For a pragmatic stance

15. Luisa Murphy
Revisiting the standard organization of multi-stakeholder initiatives (MSIs): The case of a meta-MSI in Southeast Asia

16. Friedrich Bergmann
Essays on International Trade

17. Nicholas Haagensen
European Legal Networks in Crisis: The Legal Construction of Economic Policy

18. Charlotte Biil
Samskabelse med en sommerfugle-model: Hybrid ret i forbindelse med et partnerskabsprojekt mellem 100 selvejende daginstitutioner, deres paraplyorganisation, tre kommuner og CBS

19. Andreas Dimmelmeier
The Role of Economic Ideas in Sustainable Finance: From Paradigms to Policy

20. Maibrith Kempka Jensen
Ledelse og autoritet i interaktion - En interaktionsbaseret undersøgelse af autoritet i ledelse i praksis

21. Thomas Burø
LAND OF LIGHT: Assembling the Ecology of Culture in Odsherred 2000-2018

22. Prins Marcus Valiant Lantz
Timely Emotion: The Rhetorical Framing of Strategic Decision Making

23. Thorbjørn Vittenhof Fejerskov
Fra værdi til invitationer - offentlig værdiskabelse gennem affekt, potentialitet og begivenhed

24. Lea Acre Foverskov
Demographic Change and Employment: Path dependencies and institutional logics in the European Commission

25. Anirudh Agrawal
A Doctoral Dissertation

26. Julie Marx
Households in the housing market

27. Hadar Gafni
Alternative Digital Methods of Providing Entrepreneurial Finance

28. Mathilde Hjerrild Carlsen
Ledelse af engagementer: En undersøgelse af samarbejde mellem folkeskoler og virksomheder i Danmark

29. Suen Wang
Essays on the Gendered Origins and Implications of Social Policies in the Developing World

30. Stine Hald Larsen
The Story of the Relative: A Systems-Theoretical Analysis of the Role of the Relative in Danish Eldercare Policy from 1930 to 2020

31. Christian Casper Hofma
Immersive technologies and organizational routines: When head-mounted displays meet organizational routines

32. Jonathan Feddersen
The temporal emergence of social relations: An event-based perspective of organising

33. Nageswaran Vaidyanathan
ENRICHING RETAIL CUSTOMER EXPERIENCE USING AUGMENTED REALITY

2021

1. Vanya Rusinova
The Determinants of Firms' Engagement in Corporate Social Responsibility: Evidence from Natural Experiments

2. Lívia Lopes Barakat
Knowledge management mechanisms at MNCs:
The enhancing effect of absorptive capacity and its effects on performance and innovation

3. Søren Bundgaard Brøgger
Essays on Modern Derivatives Markets

4. Martin Friis Nielsen
Consuming Memory: Towards a conceptualization of social media platforms as organizational technologies of consumption

05. Fei Liu
Emergent Technology Use in Consumer Decision Journeys: A Process-as-Propensity Approach

06. Jakob Rømer Barfod
Ledelse i militære højrisikoteams

07. Elham Shafiei Gol
Creative Crowdwork Arrangements

08. Árni Jóhan Petersen
Collective Imaginary as (Residual) Fantasy: A Case Study of the Faroese Oil Bonanza

09. Søren Bering
"Manufacturing, Forward Integration and Governance Strategy"

10. Lars Oehler
Technological Change and the Decomposition of Innovation:
Choices and Consequences for Latecomer Firm Upgrading: The Case of China's Wind Energy Sector

11. Lise Dahl Arvedsen
Leadership in interaction in a virtual context:
A study of the role of leadership processes in a complex context, and how such processes are accomplished in practice

12. Jacob Emil Jeppesen
Essays on Knowledge networks, scientific impact and new knowledge adoption

13. Kasper Ingeman Beck
Essays on Chinese State-Owned Enterprises:
Reform, Corporate Governance and Subnational Diversity

14. Sönnich Dahl Sönnichsen
Exploring the interface between public demand and private supply for implementation of circular economy principles

15. Benjamin Knox
Essays on Financial Markets and Monetary Policy

16. Anita Eskesen
Essays on Utility Regulation: Evaluating Negotiation-Based Approaches inthe Context of Danish Utility Regulation

17. Agnes Guenther
Essays on Firm Strategy and Human Capital

18. Sophie Marie Cappelen
Walking on Eggshells: The balancing act of temporal work in a setting of culinary change

19. Manar Saleh Alnamlah
About Gender Gaps in Entrepreneurial Finance

20. Kirsten Tangaa Nielsen
Essays on the Value of CEOs and Directors

21. Renée Ridgway
Re:search - the Personalised Subject vs. the Anonymous User

22. Codrina Ana Maria Lauth
IMPACT Industrial Hackathons: Findings from a longitudinal case study on short-term vs long-term IMPACT implementations from industrial hackathons within Grundfos

23. Wolf-Hendrik Uhlbach
Scientist Mobility: Essays on knowledge production and innovation

24. Tomaz Sedej
Blockchain technology and inter-organizational relationships

25. Lasse Bundgaard
Public Private Innovation Partnerships: Creating Public Value & Scaling Up Sustainable City Solutions

26. Dimitra Makri Andersen
Walking through Temporal Walls: Rethinking NGO Organizing for Sustainability through a Temporal Lens on NGO-Business Partnerships

27. Louise Fjord Kjærsgaard
Allocation of the Right to Tax Income from Digital Products and Services: A legal analysis of international tax treaty law

28. Sara Dahlman
Marginal alternativity: Organizing for sustainable investing

29. Henrik Gundelach
Performance determinants: An Investigation of the Relationship between Resources, Experience and Performance in Challenging Business Environments

30. Tom Wraight
Confronting the Developmental State: American Trade Policy in the Neoliberal Era

31. Mathias Fjællegaard Jensen
Essays on Gender and Skills in the Labour Market

32. Daniel Lundgaard
Using Social Media to Discuss Global Challenges: Case Studies of the Climate Change Debate on Twitter

33. Jonas Sveistrup Søgaard
Designs for Accounting Information Systems using Distributed Ledger Technology

34. Sarosh Asad
CEO narcissism and board composition: Implications for firm strategy and performance

35. Johann Ole Willers
Experts and Markets in Cybersecurity On Definitional Power and the Organization of Cyber Risks

36. Alexander Kronies
Opportunities and Risks in Alternative Investments

37. Niels Fuglsang
The Politics of Economic Models: An inquiry into the possibilities and limits concerning the rise of macroeconomic forecasting models and what this means for policymaking

38. David Howoldt
Policy Instruments and Policy Mixes for Innovation: Analysing Their Relation to Grand Challenges, Entrepreneurship and Innovation Capability with Natural Language Processing and Latent Variable Methods

2022

01. Ditte Thøgersen
Managing Public Innovation on the Frontline

02. Rasmus Jørgensen
Essays on Empirical Asset Pricing and Private Equity

03. Nicola Giommetti
Essays on Private Equity

04. Laila Starr
When Is Health Innovation Worth It? Essays On New Approaches To Value Creation In Health

05. Maria Krysfeldt Rasmussen
Den transformative ledelsesbyrde – etnografisk studie af en religionsinspireret ledelsesfilosofi i en dansk modevirksomhed

06. Rikke Sejer Nielsen
Mortgage Decisions of Households: Consequences for Consumption and Savings

07. Myriam Noémy Marending
Essays on development challenges of low income countries: Evidence from conflict, pest and credit

08. Selorm Agbleze
A BEHAVIORAL THEORY OF FIRM FORMALIZATION

09. Rasmus Arler Bogetoft
Rettighedshavers faktisk lidte tab i immaterialretssager: Studier af dansk ret med støtte i økonomisk teori og metode

10. Franz Maximilian Buchmann
Driving the Green Transition of the Maritime Industry through Clean Technology Adoption and Environmental Policies

11. Ivan Olav Vulchanov
The role of English as an organisational language in international workplaces

12. Anne Agerbak Bilde
TRANSFORMATIONER AF SKOLELEDELSE - en systemteoretisk analyse af hvordan betingelser for skoleledelse forandres med læring som genstand i perioden 1958-2020

13. JUAN JOSE PRICE ELTON
EFFICIENCY AND PRODUCTIVITY ANALYSIS: TWO EMPIRICAL APPLICATIONS AND A METHODOLOGICAL CONTRIBUTION

14. Catarina Pessanha Gomes
The Art of Occupying: Romanticism as Political Culture in French Prefigurative politics

15. Mark Ørberg
Fondsretten og den levende vedtægt

16. Majbritt Greve
Maersk's Role in Economic Development: A Study of Shipping and Logistics Foreign Direct Investment in Global Trade

17. Sille Julie J. Abildgaard
Doing-Being Creative: Empirical Studies of Interaction in Design Work

18. Jette Sandager
Glitter, Glamour, and the Future of (More) Girls in STEM: Gendered Formations of STEM Aspirations

19. Casper Hein Winther
Inside the innovation lab - How paradoxical tensions persist in ambidextrous organizations over time

20. Nikola Kostić
Collaborative governance of inter-organizational relationships: The effects of management controls, blockchain technology, and industry standards

21. Saila Naomi Stausholm
Maximum capital, minimum tax: Enablers and facilitators of corporate tax minimization

22. Robin Porsfelt
Seeing through Signs: On Economic Imagination and Semiotic Speculation

23. Michael Herburger
Supply chain resilience – a concept for coping with cyber risks

24. Katharina Christiane Nielsen Jeschke
Balancing safety in everyday work - A case study of construction managers' dynamic safety practices

25. Jakob Ahm Sørensen
Financial Markets with Frictions and Belief Distortions

26. Jakob Laage-Thomsen
Nudging Leviathan, Protecting Demos - A Comparative Sociology of Public Administration and Expertise in the Nordics

27. Kathrine Søs Jacobsen Cesko
Collaboration between Economic Operators in the Competition for Public Contracts: A Legal and Economic Analysis of Grey Zones between EU Public Procurement Law and EU Competition Law

28. Mette Nelund
Den nye jord – Et feltstudie af et bæredygtigt virke på Farendløse Mosteri

29. Benjamin Cedric Larsen
Governing Artificial Intelligence – Lessons from the United States and China

30. Anders Brøndum Klein
Kollektiv meningsdannelse iblandt heterogene aktører i eksperimentelle samskabelsesprocesser

31. Stefano Tripodi
Essays on Development Economicis

32. Katrine Maria Lumbye
Internationalization of European Electricity Multinationals in Times of Transition

33. Xiaochun Guo
Dynamic Roles of Digital Currency – An Exploration from Interactive Processes: Difference, Time, and Perspective

34. Louise Lindbjerg
Three Essays on Firm Innovation

35. Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

36. Hanna Nyborg Storm
Cultural institutions and attractiveness – How cultural institutions contribute to the development of regions and local communities

37. Anna-Bertha Heeris Christensen
Conflicts and Challenges in Practices of Commercializing Humans – An Ethnographic Study of Influencer Marketing Work

38. Casper Berg Lavmand Larsen
A Worker-Centered Inquiry into the Contingencies and Consequences of Worker Representation

39. Niels le Duc
The Resource Commitment of Multinational Enterprise R&D Activities

40. Esben Langager Olsen
Change management tools and change managers – Examining the simulacra of change

41. Anne Sophie Lassen
Gender in the Labor Market

42. Alison E. Holm
Corrective corporate responses to accusations of misconduct on societal issues

43. Chenyan Lyu
Carbon Pricing, Renewable Energy, and Clean Growth – A Market Perspective

44. Alina Grecu
UNPACKING MULTI-LEVEL OFFSHORING CONSEQUENCES: Hiring Wages, Onshore Performance, and Public Sentiment

45. Alexandra Lüth
Offshore Energy Hubs as an Emerging Concept – Sector Integration at Sea

2023

01. Cheryl Basil Sequeira
Port Business Development – Digitalisation of Port Authroity and Hybrid Governance Model

02. Mette Suder Franck
Empirical Essays on Technology Supported Learning – Studies of Danish Higher Education

03. Søren Lund Frandsen
States and Experts – Assembling Expertise for Climate Change and Pandemics

04. Guowei Dong
Innovation and Internationalization – Evidence from Chinese Manufacturing Enterprises

05. Eileen Murphy
In Service to Security – Constructing the Authority to Manage European Border Data Infrastructures

06. Bontu Lucie Guschke
THE PERSISTENCE OF SEXISM AND RACISM AT UNIVERSITIES – Exploring the imperceptibility and unspeakability of workplace harassment and discrimination in academia

07. Christoph Viebig
Learning Entrepreneurship – How capabilities shape learning from experience, reflection, and action

08. Kasper Regenburg
Financial Risks of Private Firms

09. Kathrine Møller Solgaard
Who to hire? – A situated study of employee selection as routine, practice, and process

10. Jack Kværnø-Jones
Intersections between FinTech Imaginaries and Traditional Banking – A study of disciplinary, implementary, and parasitic work in the Danish financial sector

11. Stine Quorning
Managing Climate Change Like a Central Banker – The Political Economy of Greening the Monetary Technocracy

12. Amanda Bille
No business without politics – Investigating the political nature of supply chain management

13. Theis Ingerslev Jensen
Essays on Empirical Asset Pricing

14. Ann Fugl-Meyer
The Agile Imperative – A Qualitative Study of a Translation Process in the Danish Tax Administration

15. Nicolai Søgaard Laursen
Longevity risk in reinsurance and equity markets

16. Shelter Selorm Kwesi Teyi
STRATEGIC ENTREPRENEURSHIP IN THE INFORMAL ECONOMY

17. Luisa Hedler
Time, Law and Tech – The introduction of algorithms to courts of law

18. Tróndur Møller Sandoy
Essays on the Economics of Education

19. Nathan Rietzler
Crowdsourcing Processes and Performance Outcomes

20. Sigrid Alexandra Koob
Essays on Democracy, Redistribution, and Inequality

21. David Pinkus
Pension Fund Investment: Implications for the Real Economy

22. Sina Smid
Inequality and Redistribution – Essays on Local Elections, Gender and Corruption in Developing Countries

23. Andreas Brøgger
Financial Economics with Preferences and Frictions

24. Timothy Charlton-Czaplicki
Arendt in the platformised world – Labour, work and action on digital platforms

25. Letícia Vedolin Sebastião
Mindfulness and Consumption: Routes Toward Consumer Self-Control

26. Lotte List
Crisis Sovereignty – The Philosophy of History of the Exception

27. Jeanette Walldorf
Essays on the Economics of Education and Labour Market

28. Juan Camilo Giraldo-Mora
It is Along Ways – Global Payment Infrastructure in Movement

29. Niels Buus Lassen
THE PREDICTIVE POWER OF SOCIAL MEDIA DATA

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens