CBS

COPENHAGEN BUSINESS SCHOOL

HANDELSHØJSKOLEN

SOCIAL MEDIA & THE DANISH STOCK MARKET

BY PIL NYROP SKØTT,
MSc. International Business, Copenhagen Business School

Student number: 51622
Contract number: 11613
Supervisor: Abid Hussain
Hand-in date: May 15, 2018
Number of characters: 175,860

Abstract

Background – Based on findings mainly in the American stock market, it has been suggested that social media influence stock markets in a way, which may be used to maximize profit for investors. No such data are available for the Danish investors and stock market. Therefore, the present study is aimed at elucidating this question.

Methods – The awareness of, interest in, desire to use, as well as the actual use of social media trading[1] of four Danish professional investors are investigated through semi-structured interviews and the same are examined for 98 private Danish investors in a survey using a questionnaire validated through a pilot study. The relationship between social media and stock performance is examined by running timed comparisons of variations in activity of various social media and a set of stock performance indicators from five Danish companies listed in the C20 index. Social media activity is cumulated in 3-months periods from 2nd quarter 2015 to 4th quarter 2017. The stock performance is determined in the same quarters by comparing the change from day 1 in the quarter to first day in the next quarter. In one set of comparisons the stock performance is tested in the next quarter and compared to social media activity in the previous quarter. The covariation between social media activity and stock performance indicators are tested by simple and multiple regression analysis (STATA). For statistical analysis the Bonferroni correction and Benjamini-Hockberg False Discovery Rate correction methods are applied to correct for multiple comparisons.

Results – The professional investors are generally aware of social media trading. Two of these investors have shown a further interest, but none actually have a desire to use social media trading. For the private investors, less than half are aware of social media trading, nine have a desire to use it and three indicate that they already use it. The analyses of correlations between social media activity and stock performance show significant correlations in 11 cases. Nine of these are related to the company Genmab and all correlate social media activity and stock performance within the same quarters. The two others relate to Danske Bank and Bavarian Nordic, and these correlate social media activity in one quarter to stock performance in the next quarter. The social media parameters that have influence on stock performance are: mentions on portals, positive mentions on portals, number of posts, number of articles, number of tweets, negative tweets and number of positive sentiments published on portals. In most cases, it has no influence on stock performance whether the general sentiments of the social media activity are positive or negative. In 9 out of the 11 significant models, the stock performance indicator is

[1] Social media trading is a stock trading strategy that uses social media content volume or social media content sentiments to generate buy and sell signals for any stock exchange. Social media trading is described more in depth in the Conceptual Framework.

the P/E ratio. It is noticeable that mentions on Facebook, the most used social media in Denmark, have no influence on stock performance.

Conclusions – Social media trading has not been utilized by Danish investors so far. Even though 11 significant correlations originate from the quantitative part of this study, a solid base for implementation of social media trading has not yet been established in the Danish market.

Keywords and phrases – social media trading, stock market predictions, social media, Denmark, Danish stocks, regression analysis, predictive power, Bonferroni correction, Benjamini-Hockberg correction.

Foreword

To Michele Colombo: I am grateful for the help and guidance you provided to me at CBS two full weekends in January. My own skills of extracting and cleaning very unstructured data at that point of time was not enough for the dataset downloaded from SentiOne. However, with several hours of your help, I gained so much knowledge on how to deal with the 27 million data points. This project would not have been possible without your help.

To Mr. Bersant Hobdori: You have been such a big help in the confirmation of the regression models. Without your guidance and persistent help, this paper's quality would not have been the same.

To my supervisor Abid Hussain: Thanks for keeping me on track and for giving professional and clarifying supervision!

To Søren Hansen from CBS's library: Thanks for the help with the data collection of the stock performance indicators.

Thanks also to the four professional investors who participated in the interviews.

TABLE OF CONTENT

TABLE OF FIGURES

1. Introduction

The term '*invest'* origins from the Latin word '*investire',* which means 'to cloth' (Online Etymology Dictionary). Investire related to financial activities became a part of the English vocabulary from the Italian language in the early 17th century (Merriam-Webster, 2017). It was first used in the first half of the 15th century in North Italy. Since then, different perspectives of the phenomenon have emerged. However, the purpose of *investing* has always been to obtain an additional income, profit or capital on the initial investment (Caroe, 2016). As of today, the most common kinds of investments are *bonds, stocks* and *mutual funds* (Ray, 2000). A bond is a loan that the buyer of the bond offers to the issuer of the bond, which would typically be governments or companies. Bonds have a date of maturity. If investors want to access or sell their bonds prior to the maturity date, it will have consequences for the investor such as additional fees as well as the value of the bond at the time of selling. Buying stocks is the same as buying a share of a company, making the investor an owner of parts of the company. Only companies issue stocks. Returns on stocks can be made in two ways: firstly, by receiving dividends if the company pays out dividends. Secondly, by selling the stock at a higher price than the buying price. Warren Buffet is one of such recognized investors who has been able to make remarkable returns by buying stocks cheaply and selling at a higher price (Business Insider, 2018; Forbes, 2018), which clearly indicates the economic benefits of having the right market timing. Mutual funds are portfolios of stocks and bonds managed and composed by investment professionals.

This study focuses exclusively on stocks. Stocks are traded on stock market exchanges such as New York Stock Exchange, NASDAQ OMX and AMEX. The stock investors include both private investors[2] and professional investors[3] (Infront, 2018). Especially, the professional investors are constantly seizing new investment opportunities and new *trading strategies* (Odean, 1999). A trading strategy describes and guides when to enter the trade, when to exit and how the money should be managed, and is in some cases expressed mathematically. Numerous methods are applied to accomplish a trading strategy. According to Fung and Hsieh (1997), investors should always be of the limitations of the trading strategy and the risk related to the strategy. Examples of stock trading strategies are: *Day Trading, Slope Performance Trend* and *Social Trading* (StockCharts, 2018). Most of

[2] A private investor is a person who does not perform and is not in any other way engaged in giving investment services. A private investor exclusively invests his/her own wealth (Infront, 2018).

[3] A professional investor can be a corporation, a partnership, proprietorship or any other entity that run investment or banking services. A professional investor can also be a person that is involved in giving investment services. The definition of giving investment services is: *"Being authorized to give investment services or registered with SEC (or authorities with similar functions in other countries)"* (Infront, 2018).

these trading strategies have been known for decades. However, new strategies such as social listening based strategies and social trading have emerged with the development and spread of social media over the past 10 years (Chaffey, 2017). A more recent development of a trading strategy is based on the content posted on social media. For instance, the financial Big Data analytics company Market Prophit launched in May 2015 the first social media sentiment stock market index allowing investors to get the pulse of the crowd sentiment (Market Prophit, 2015). Due to the increased number of firms that have introduced similar indices or eventually algorithms to explore the economic benefits that social media insights may give to financial professionals, it has become clear that these types of trading strategies are of increasing interest. More and more academics of both the economic, financial, technological, - and psychological research fields have explored this new investment behavior (Schoen et al., 2013; Nguyen, Shirai and Velcin, 2015; Wolfers and Zitzewitz, 2004; Asur and Huberman, 2010; Bordino et al., 2012). Social media analysis is gaining more and more acceptance for its ability to e.g. forecast sales, predicting the outcome of an election or the development in TV reality shows. However, the acceptance of social media to predict financial markets is very limited (Kooijman, 2014) even though, lately, statements on social media have shown to have financial impacts. Some examples are the plummeting Snap Inc. stock after the celebrity Kylie Jenner tweeted that she was done with SnapChat (The Verge, 2018; Bloomberg, 2018; CNN, 2018) and the plummeting of several huge, listed firms including Toyota (Mediairite, 2017), Amazon (Deadline, 2018) and the raise of the value of Boing (CNBC, 2016) as a consequence of tweets published by the President of the U.S., Donald Trump about those firms. Recent examples show that totally unknown persons' statements about firms with enough support on social media from many other unknown people can also impact the value of a company. One of such examples is the United Airlines stock that dropped $1.4 Billion after a controversial removal of a passenger from an overbooked airplane (Fortune, 2017).

Although there are plenty of examples, no uniform term covers the phenomenon of stock trading on the basis of what is written on social media. It can be due to insufficient research on the topic, insignificant results limiting the trust or just because it is a new field of research. Existing literature covers mainly American stocks or indices, whereas no research has ever been done on i.e. DAX or London Stock Exchange or NASDAQ OMX. The author of this study wants to reduce the gap by conducting a study on the Danish stock exchange, NASDAQ OMX.

1.1. *Definition of research problem and research questions*

"*A research problem is a clear expression about an area of concern (…) or within existing practice that points to a need for meaningful understanding and deliberate investigation. A research problem does not state how to do something*" (Bryman, 2007, p. 5). This paper aims to answer the following research problem:

Do Danish investors use social media trading[4], and does a relationship between social media mentions and Danish listed companies' stock performances exist?

The research problem of the thesis will be answered by investigating the following research questions:

I. *To what extent do social media influence the decision-making process for Danish professional investors?*

II. *To what extent do social media influence the decision-making process for Danish private investors?*

III. *Does social media activity about selected Danish C20 companies influence their short-term stock performance (i.e. within the same quarter)?*

IV. *Does social media activity about selected Danish C20 companies influence their long-term stock performance?*

This proposition seeks to be investigated using interviews with professional investors to answer research question 1, questionnaires to private investors to answer research question 2, Stock Performance Data retrieved from DataStream[5] and Social Media Data retrieved from SentiOne[6] to answer research question 3 and research question 4.

1.2. *Delimitations*

The research questions of this thesis limit the study by only focusing on Danish stocks, however, the findings may be applicable to foreign stocks as well. Furthermore, this study is only investigating Danish investors' degree of use of social media trading. However, due to the free market, foreigners

[4] Social media trading is a stock trading strategy that uses social media content volume or social media content sentiments to generate buy and sell signals for any stock exchange. Social media trading is described more in depth in the Conceptual Framework.

[5] DataStream is a global financial macroeconomic data platform from Thomson Reuters. It provides over 10 million economic time series on equities, stock market indices, currencies, company fundamentals, fixed income securities and key economic indicators for 162 markets (EUI, 2018; Thomson Reuters, 2018).

[6] SentiOne is a social listening tool that is capable of monitoring both statements and articles from the whole internet. It covers social media like Facebook, Instagram, Twitter, Google+, Youtube and more, and also blogs, forums and online portals. Both historical social media data can be collected through SentiOne, but also real-time data can be collected (SentiOne, 2018; Crunchbase, 2018)

are also able to trade the Danish stocks. Foreign traders' usage of social media trading is not examined in this study.

Moreover, the study is confined to an analysis of five randomly picked stocks on the Danish C20 stock index. The number of social media sources included for this study is high, but it is likely that other factors than social media affect the stock performance, and these are not investigated. Another risk is that both the social media variables *and* the stock performance variables are affected by the same or correlated exogenous factors. It is not investigated neither. The information used and the data collected for this part of the project are quantitative data. Research such as Science (2001) criticizes the nature of this kind of data due to its lack of ability to capture derived details. However, big data analysis is particularly well suited to discover peculiar details of a dataset (Mayer-Schönberger and Cukier, 2013).

1.3. *Motivation*

The motivation for choosing this topic originates from the researcher's interest in Big Data and the financial sector. Furthermore, the researcher wanted to learn more about how Big Data could be collected, processed and used. Moreover, existing literature seems to lack information on whether social media impact the performance of Danish stocks. If the researcher could demonstrate that social media actually do influence stock performance, it would be a significant contribution to both existing literature and practice.

1.4. *Structure of the report*

The study will first review existing academic literature in the fields relevant for the research questions. It will then set the frame of the study by explaining the conceptual framework that defines and describes concepts and approaches that are crucial for the reader to understand. After the conceptual framework is explained, the theoretical framework of the AIDA model (Awareness, Interest, Desire, Action) is provided. After the theoretical framework has been explained, the study will proceed with focus on the methodology to collect both primary and secondary data. Subsequently, the collected data will be analyzed using the AIDA model. Finally, the results will be discussed and conclusions drawn. Figure 1 outlines the structure of the thesis.

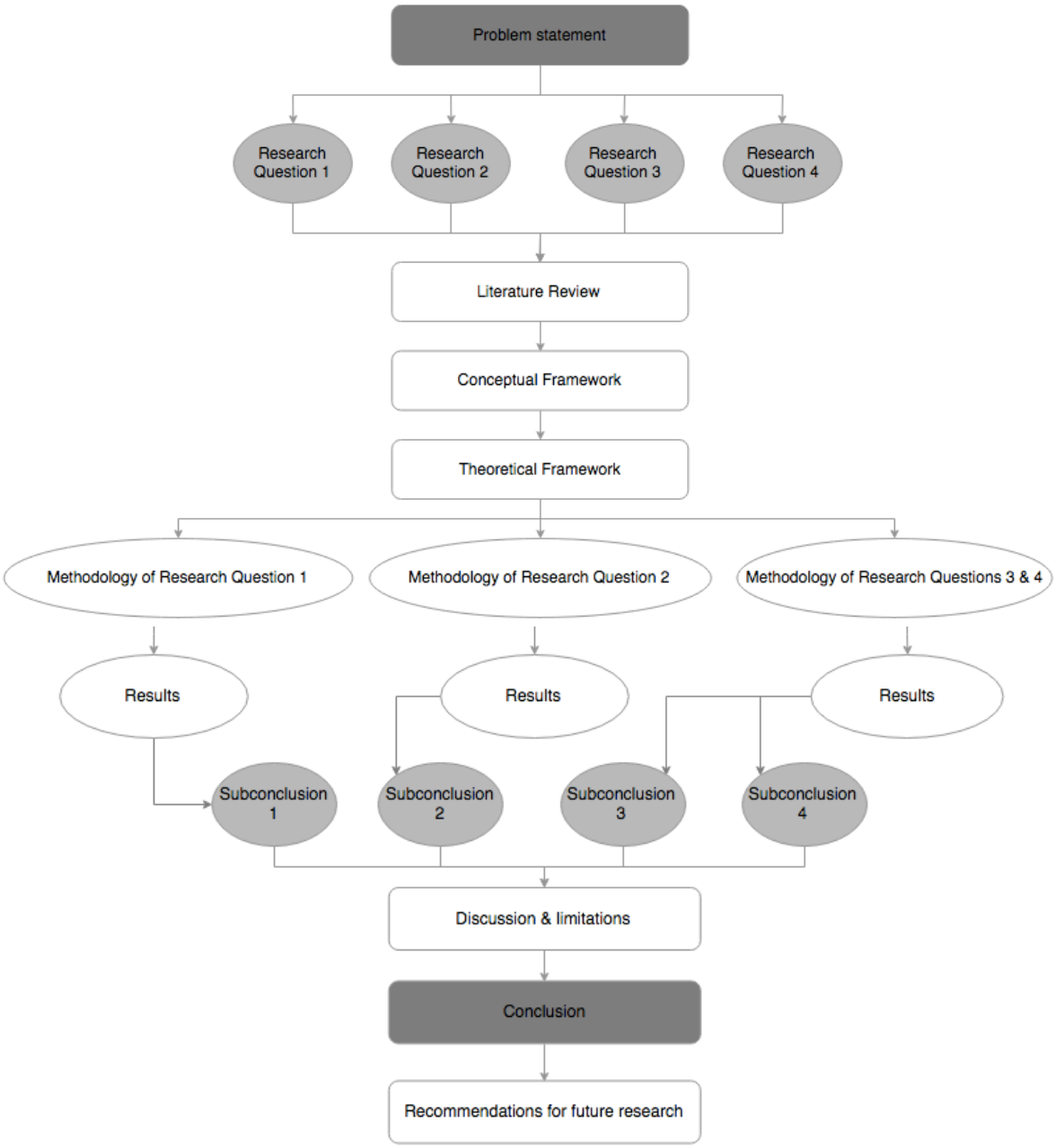

Figure 1: Structure of the Thesis

2. Literature review

According to the Gutman Library at Harvard Business School, a literature review is *"an assessment of a body of research that addresses a research question"* (Gutman Library at Harvard Business School, line 7). The study will review literature about investors' behavior on the stock market and how they make decisions. In addition, it will review literature on the use of social media to predict the stock market movements.

2.1. *Understanding investor behavior and decisions*

In the recent decades, it has become increasingly popular by professional investors and private investors to invest in financial markets (Bikas et al., 2013). Existing literature such as Lakonoshok, Shleifer and Vishny (1992), Sias (2004) and Zhang (2013) differentiate between investors depending on their investment horizon. It can either be a *long-term perspective* or a *short-term perspective*. The investment horizon refers to the duration of time that an investor expects to hold a stock. Scholars such as Boudoukh and Richardson (1993) and Barber and Lyon (1997) claim that stock investments are optimal for long-term investors, but Ryu's (2012) findings suggest that day trading of stocks can be very profitable if high quality data are accessible.

Fully independent of the investment horizon, most researchers agree that investors need some kind of indicator to decide when to make their investment decision. Such an indicator is called a *stock market indicator* and is defined as *"a ratio or a formula that explains gains and losses in stocks"* (MarketSmith, 2018). Stock market indicators are used to predict future returns in financial assets (Nazário et al. 2017) by studying historical market data, mainly price and volume (Park and Irwin, 2007; Wei, Chen and Ho, 2011), or price-to-book ratio[7] (Boykin, 2017), and price-to-earnings ratio[8] (Rahman, 2011; Shamsudin, Mahmood and Ismail, 2013).

Traditional finance theories assume investors to be *rational* (Simon, 1955). A rational investor is defined as an investor who makes decisions based on all available information (Simon, 1955). Traditional theories such as efficient market hypothesis (EMH) theory and modern portfolio theory are created under the assumption that all individuals are behaving rationally. However, in 1955 Simon (1955) questions human beings' rationality when making decisions. The concept is captured by the term "*bounded rationality*". Simon (2000) defines bounded rationality as *"rational choice that takes into account the cognitive limitations of the decision maker—limitations of both knowledge and computational capacity"* (p. 291). The definition implies human beings' ability to process information and on that basis make sub-optimal decisions. The question of whether investors behave rationally has caused much debate in the financial literature over the years (Wang, 2015; Hirshleifer, 2007). Babajide and Adetiloye (2012) and Bashir et al. (2013) state that investors often overreact to market information, because many investors act upon intuition and forget the general principles of investment theory (De Bondt, 1998). Also, academics of the behavioural finance field indicate that human beings do not behave as rationally as EMH economists suggest (Bakar and Yi, 2016). Scholars in behavioural

[7] The price-to-book ratio measures a company's market value relative to its book value (Sharma et al., 2013).
[8] The price-to-earnings ratio measures a company's share price relative to its per-share earnings (Fun & Basana, 2012).

finance emphasize that psychological factors such as emotions and cognitive errors impact the behaviour of individuals and groups including professional investors such as bankers, strategists and portfolio managers (Bakar and Yi, 2016; Kengatharan 2014). These psychological aspects that affect investors' decision making in the stock market include among others: overconfidence bias, conservatism bias, herding and availability bias, and can lead to market inefficiencies. Both Benjamin Graham and Warren Buffet, most influential investors in history, acknowledge these phenomena.

Scholars have preciously applied *Theory of Planned Behaviour* (Ajzen, 1991) in their desire to understand investment behaviour. Warsame and Ireri (2016), East (1993) and Sondari and Sudarsono (2015) have applied Theory of Planned Behavior to study the reasoning behind investors' investment decisions by understanding *attitude*, *subjective norm*, *perceived behavioural control* and *behavioural intention* and *actual behaviour*. Recently, an increasing number of scholars find it relevant to study peoples' *awareness, interest, desire* and *actions/purchases* when they investigate who and why people are becoming motivated to act on purchase (Barry and Howard, 1990). The model is called the *AIDA model*. The model is invented by Elmo St. Lewis in 1898 and it explains four cognitive steps of an individual's experience from being aware of a new idea or product to the idea is executed or the product is purchased (Michaelson and Stacks, 2011). Over time the model has been modified to fit technological developments (Ashcroft and Hoey, 2001; Lassen, Madsen and Vatrapu, 2014).

2.2. *Definition and few essentials on social media*

Social media is defined as "*a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content*" (cited in Jussila et al., 2013 p. 2).

The interactions between users distinguish social media from traditional online media. Whereas traditional types of media are most often one-way communication forms enabling the business to send messages to the customers/potential customers, social media applications encourage users to share their experiences, opinions and knowledge, and thus, facilitate collaboration and two-way communication methods. Cooke and Buckley (2008) state that social media include: wikis (e.g. Wikipedia), blogs and micro-blogs (e.g. Twitter), social networking sites (e.g. Facebook) and social content communities (e.g. Instagram and Youtube). The functionality of these platforms vary and so do the rules of utilization and its functions (Jussila et al., 2013), which imply that various platforms can be used with different functionalities, and that one person is allowed to use more than one

platform, and even to publish similar content on all platforms. It provides researchers and practitioners with great opportunities to know their audience much better: *"when the data is collected passively while people do what they normally do anyway, the old biases associated with sampling … disappear. We can now collect information that we couldn't before"* (Mayer-Schönberger and Cukier, 2013, p. 30).

Social media has experienced an explosive increase in the number of users over the past few years (Statista, 2018; Hanna, Rohm and Crittenden, 2011). As of end of January 2018, there are 2.13 billion monthly active Facebook users, 1.5 billion on Youtube, 800 million users of Instagram and 330 million monthly active users of Twitter (DMR, 2018). In Denmark 80% of the population is active on Facebook, 55% of the population has a Youtube account, 29% has an Instagram account and 17% has a Twitter account. As of 2017, 57% of the Danish population actively uses between two and more than six different social media (Bureau, 2017).

2.3. *Social media's impact on the stock market*

Already before the introduction of social media, Bagnoli et al. (1999) saw the potential in social networks in the way it could become a useful source of valuable information in a context where earnings forecasts among traders could be exchanged.

Since the emergence of social media, a growing number of researchers examine the application of social media in relation to trading in various contexts and with various methodologies. Some scholars study the correlation between stock market variables and a micro-blogging variable i.e. the sentiment of the content posted (Bissattini and Christodoulou, 2013), whereas others investigate how posting volume on social media affects a stock price (Tumarkin and Whitelaw, 2001) and how two or more micro-blogging variables such as the tweet sentiment and the message volume together can predict the stock market (Sprenger and Welpe, 2010).

Studying sentiment analysis became increasingly interesting for scholars recently, as the subjective pieces of texts are found useful for various applications (Nanli et al., 2012). However, most work done in academic fields mainly focus on "*framework and lexicon construction, feature creation and polarity determination*" (Nanli et al., 2012, p. 1). Conversely, the following review of literature studied previous research that applied sentiment analysis to study the financial impact of sentiments retrieved from social media. Bollen et al. (2010) use Tweet sentiments (both the polarity of positive/negative and a six sentiment dimensions) to predict the Dow Jones Industrial Average with 86 % accuracy by using sentiment analysis. These revealing results are confirmed by Tayal and Satya (2009) and Oh and Sheng (2011). For instance, Oh and Sheng (2011)'s research indicates that

the sentiment of micro-blog content about stocks can predict market returns. In addition, Ranco et al. (2015) state that Tweet sentiments can indicate the direction of abnormal returns. Shen et al. (2017) suggest that the search frequency on social media of a stock in Baidu Index was a proxy for the investor's degree of attention towards a stock. The more attention a company attracts on social media, the higher the dissemination of information about the company, the more purchase actions are initiated, and, therefore, the higher the market efficiency. To complete the study of Bollen et al. (2010), they create a sentiment index, which subsequently has been used by other scholars including Chyan and Lengerich (2011). In the study, it is emphasized that stock-chasing agents that use Bollen et al. (2010)'s sentiment index have higher accuracy in the stock predictions compared to the stock-chasing agents not using the index. However, the stock-chasing agents that use the index have lower yields than the ones not using it due to extreme losses when wrong predictions occur (Chyan and Lengerich, 2011). Similar to Bollen et al. (2010), Sul, Dennis and Yuan (2016) collect data from public tweets about S&P500 firms. In opposition to Bollen et al. (2010), Sul, Dennis and Yuan (2016) use a social media metric[9]. The social media metric is used as the predictor and the findings suggest that the social media metric is more strongly correlated with firm stock performance than conventional media. Bissattini and Christodoulou (2013) and Ruiz et al. (2012) both use linear regression to predict returns and stock market movements based on micro-blogs sentiments, which can be used to develop and implement trading strategies (Bissattini and Christodoulou, 2013).

Li et al. (2014) suggest that investors' emotions can be influenced by public sentiments. Coupled with the literary evidence provided by Tetlock (2011), it is argued that investment decisions are impacted by public sentiments. Tetlock (2011) shows that media pessimism can predict down movements in stock prices. Contrary to Li et al. (2014) and Tetlock (2011), Oliveira et al. (2013) conclude that sentiment variables have no predictive power over returns of investments. Further evidence against sentiment analysis as a tool to predict stock behaviour may lie in the findings of Logunov and Panchenko (2011). They do not find a relationship between the sentiment index developed by Bollen et al. (2010) and the Dow Jones index. Lognunov and Pachenko (2011) propose that the sentiment index is too simple. Dahiru (2008) claims that some reports with insignificant findings show "*a strong tendency to accentuate the positive findings"* (p. 2) even when these were non-significant (or non-existing).

[9] A social media metric is used to gauge social media activity's impact on a company's financial performance, such as the stock performance (Misirlis & Vlachopoulou, 2018).

The research mentioned above uses sentiment analysis, which builds on semantic analysis that is dependent on accuracy of the technology. However, sentiment analysis has limitations due to lack of training of the sentiment categorizer or a high level of subjectivity. As a consequence, an increasing number of scholars, including Evangelopoulos, Magro and Sidorova (2012); Sprenger and Welpe (2010) have investigated the correlation between micro-blog network and stock market using aggregate data[10] of micro-blogs. Evangelopoulos, Magro and Sidorova (2012) have examined an aggregate of Tweets to predict future stock prices of 18 of the largest publicly traded companies on the US stock exchange. Evangelopoulos, Magro and Sidorova (2012) have found the predictor to be valid. Evangelopoulos, Magro and Sidorova (2012) have concluded that their findings capture both rational decision making dimensions and emotional decision making dimensions such as animal spirit[11].

2.4. *Criticism of the usage of Social Media Data*

In spite of the potential advantages of social media, substantial pessimism surrounds Social Media Data and Social Media Data analysis (Ruth and Pfeffer, 2014). The criticism of social media and Social Media Data analysis as a tool to predict stock behaviour have been addressed by numerous academics from various fields including Phillips et al. (2017), Nanli et al. (2012) and Beigi et al. (2016). Given the importance of these issues to the present study (see section *Secondary data*), some of the precautions will be described in the following.

Phillips et al. (2017, p. 1) have argued that social media platforms provide '*users with an online identity*", and thus, it is argued that an online identity not necessarily has to be a *real identity*. Consequently, it is harder to track the person behind an 'online identity' profile's motivation for posting and sharing content. Along similar lines Waddell (2018) have argued that the escalation in use of robots to push out information on social media challenges the credibility of social media because it is difficult to distinguish between intelligent robots and human beings when analysing social media content. Another risk Phillips et al. (2017, p. 1) outline is that "*..this information can be collected and mined by virtually anyone who wishes to use it*". However, the easy access to large amounts of data can according to recent research on Social Media Data be abused (i.e. Ratkiwicz et al., 2011). In a perfect market, all publicly available information, including relevant information derived from social media, should, in principle, be used to set the market price of the stocks. On the other hand, the recent scandals,

[10] Aggregate data are data collected from multiple sources and/or on multiple measures and expressed in a summary form, most often, for the purpose of statistical analysis (Rafanelli, 1995).
[11] "*Animal spirit is a motivational force within individuals that moves them toward being restless or inconsistent in the face of uncertain economic factors*" (Akerlof & Shiller, 2009, p. 11)

e.g. Cambridge Analytica and Facebook, indicate that these information ecosystems are not mature, and that the legal and political framework around the use of these information sources is not yet perfect.

Furthermore, due to the nature of Social Media Data being very noisy and of mixed data quality (Phillips et al., 2017), it requires heavy data cleaning and an ability to handle and store massive quantities of data.

Researchers have questioned the effectiveness of Social Media Data driven models (Matthews et al., 2018). Other problems relate to the monitoring and the analysis of sentiments or the challenge of omitted social media content. Some firms delete negative comments on their Facebook wall, or even pay the negative authors to delete their comment. This distortion is hard to monitor and requires an immediate data collection to overcome this problem.

Nanli et al. (2012) have highlighted in their comprehensive survey of sentiment analysis that less research examine the practical challenges of using sentiment analysis on Big Social Data. However, in relation to practitioners' reviews on the implications of sentiment analysis, the Internet is exploding with criticism, which is important to assess as well. The following part presents a short review on the criticism, presented by scholars and practitioners.

The machinery performing the sentiment analysis is programmed to browse a piece of text for certain keywords it uses as proxies for the type of sentiment. Positive keywords are for instance *like, love, enjoy* and negative keywords include i.e. *hate, angry, upset* (Pang, Lee and Vaithyanathan, 2002). According to Christopher Penn[12] (cited in Shift Communications, 2015), the most challenging part of automated sentiment analysis is to understand sarcasm and irony. A post like "*Oh, yeah, Fast Food Restaurant. I just LOVE the 30 minute wait for my food"* will most often be wrongly categorized as a positive sentiment by the machinery, even though humans understand the sarcasm, and thus, categorize the sentiment as being negative. Moreover, as sentiment analysis technologies assume that sentences only contain sentiments about a single entity (Gandomi and Haider, 2015), more complex constructions of the sentiments of those sentences can be wrongly classified. Beigi, et al. (2016) highlight that the use of NPL[13] processing is difficult to do correctly without the context.

[12] Christopher Penn is a marketing keynote speaker, who has taught several executives of major venture capital firms on marketing. He has also been featured in books, newspapers such as the Wall Street Journal and the New York Times, magazines such as BusinessWeek, television, and publications.

[13] "*Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies"* (Liddy, 2001, p. 1). By using NLP, it is possible for computer to e.g. read text and measure the sentiments (SAS, 2018).

Francesco D'Orazio states: "*The main problem being that the sentiment of a sentence only rarely lies in the sentence itself and is instead rooted in the cultural context around that sentence"* (Burn-Murdock, 2013).

Franscesco D'Orazio[14] (cited in Burn-Murdock, 2013) and Kessler (2014) claim that by combining a piece of text with the author's age increases the ability of the machines to categorize sentiments more precisely. Nick Halstead[15] argues that no machines today are able to group sentiments correctly with more than 70% accuracy (Burn-Murdock, 2013; Srihari, 2015), but the technology improves constantly. As of 2015 almost all major firms use sentiment analysis (Shift Communications, 2015) to look for insights into i.e. consumer behavior and consumer satisfaction (Kennedy, 2012).

2.5. *Summary of the literature review*

No research on the relationship between social media and the Danish stock market has been conducted before. The examination of existing literature demonstrates a gap in understanding the impact of social media on Danish stocks and in understanding if social media impact investors' trade decisions on Danish stocks.

Scholars have used various models in their desire to understand investors' behaviour on the financial markets but recently a growing number of researchers using Big Social Data in their studies have applied the AIDA model.

The literature reviewed clearly indicates different findings within the research on the relationship between social media and stock performance: the majority of the researchers have found a relationship, but a minority of the literature does not report any relations. Furthermore, academic papers in the field have been written complexly, which makes such literature only readable for the few with expertise in big data analytics and financial modelling.

Social Media Data are very noisy and the data quality is mixed, which requires heavy data cleaning. It is time consuming and requires a certain sets of skills to perform the data cleaning or an advanced developed software program that can handle the quantity of data. Moreover, it is a challenge to include all historical data as people sometimes delete publications on social media, whereby an analysis can be misleading.

Sentiment analysis technologies are facing some challenges including its lacking capabilities to understand sarcasm and irony. Literature about social media includes some points of criticism, which relate to the nature of social media such as the difficulty to verify the posting person's identity

[14] Franscesco D'Orazio is CIO at FACE, facegroup.com and Pulsar
[15] CTO in Datasift

and the difficulty in determining whether the one that created content on social media is a human being or a robot.

Recently, severe abuse of Social Media Data, violations of social media users' privacy, and extensive use of robots have been hot topics.

In addition, the CTO of Datasift states that the accuracy of classification of sentiment is no more than 70%. Various scholars have studied the validity of polarity methods, and the results suggest that the degree of validity varies greatly from one dataset to another.

3. Conceptual framework

"*The conceptual framework sets the stage*" (McGaghie, Bordage and Shea, 2001), and "*it is a theory, model or approach that situates the study questions within a theoretical context*" (Beckman and Cook, 2007). This conceptual framework provides explanations and definitions on concepts that are crucial for the reader to be familiar with in order to understand both the research questions and the project as a whole. The different concepts described in the following are related to: *Concepts to help understand the research questions* and *Types of data, data collection concepts, data analytics techniques and models.*

3.1. *Concepts to help understand the research questions*

This subsection provides an explanation of the term *social media trading*, which is coined for this specific study. In addition, the *efficient market hypothesis* and *market inefficiencies and anomalies* are explained.

3.1.1. *Social media trading*

The construct '*social media trading*' is invented for this study. It is a stock trading strategy that studies social media content volume or social media content sentiments to generate buy and sell signals for any stock exchange. The present study only investigates social media trading in relation to NASDAQ OMX Copenhagen. The underlying belief in this trading strategy is that social media impact investors' stock picking decision and hereby the stock market.

3.1.2. *Efficient Market Hypothesis*

The Efficient Market Hypothesis (*EMH*) is a traditional financial theory. It states that it is impossible to beat the market (Adams et al., 2007). Furthermore, according to the theory, an organization's stock price fully reflects all relevant information available about the organization, and investors have equal access to the same information (Adams et al., 2007). Hence, stocks always trade at their fair value; making it impossible for investors to excess return, hereby beating the market. EMH assumes that investors are rational decision makers. As a result, neither feelings nor psychological decision-making can drive the market due to the unemotional nature of investors (Malkiel, 1989). Change in stock markets can only derive from general market demand, insider knowledge or from releases of news. According to the hypothesis, investors can only obtain higher returns by obtaining a riskier investment behavior. Organizational performance can, thus, be analyzed by the stock performance movements, and by an investigation of the development of historical stock prices over time. It is therefore possible to determine social media's effect on the overall value of the stock performance (Pineiro-Chousa, et al., 2017).

3.1.3. *Market inefficiencies and anomalies*

Stock market inefficiencies are distortions in the market due to unfair competition, lack of market transparency and regulatory actions (Montier, 2009). These inefficiencies and distortions create the possibility of making abnormal stock returns (Adams et al., 2007). Documented anomalies in developed economies include *seasonality,* the *turn-of-the-year-effect* and *inside information* (Montier, 2009). Seasonality is a pattern in stock price or in the variability, which regularly occurs on specific dates and times. Studies have detected seasonality in intraday, weekly, monthly and annual return data (Cross, 1973; French 1980; Gibbons and Hess, 1981; Keim and Stambaugh, 1984). The *turn-of-the-year-effect* describes a pattern where returns tend to be higher in January than the rest of the year (Ritter and Chopra, 1989). The *inside information* refers to the situation where an individual with inside information can earn excess returns (Seyhun, 1986).

3.2. *Types of data, data collection concepts, data analytics techniques and models*

This subsection provides an overview of concepts that are crucial for the understanding of the data collection, the data analytics methods and the techniques. The structure of the subsection follows J.P. Morgan (2017, p. 21)'s stepwise Big Data approach for investment professionals. The steps are: *identifying and acquiring data;* and *store, structure and pre-process data;* and *analyze data;* and *trade ideas*. The four steps are relevant to follow when understanding how data are collected and which analytical methods and techniques that will be used (J. P. Morgan, 2017). The present study finds it relevant to explain a number of important concepts, i.e. *Big Data and Big Social Data, Filtering, Time adjustments* and *Visualization, Sentiment Analysis* and *Regression Analysis* under different steps in the process. The concepts that are explained in this section will increase the understanding of the Methodology of the study.

3.2.1. *Identifying and Acquiring Data*

The first step in J. P. Morgan's (2017) four-step process is the identification and acquisition of data. This step involves identification of *where* to find the *right* data and *how* to collect the data. To understand this study's identification and acquisition of data, the reader must first understand what *Big Data* are and what *Big Social Data* are.

3.2.1.1. Big Data and Big Social Data

According to Laney (2001): '*Big Data can be defined based on large volumes of extensively varied data that are generated, captured, and processed at high velocity*' (p. 1). Value, which is the on-going assessment of the reliability of the gathered data, has recently emerged as a fourth criterion to the definition (Hitzler and Janowicz, 2013). Consequently, Big Data are the production of coherent and qualified data sets derived from many different sources (e.g stock performance sources and social media sources). According to Haughton (2013), Big Data typically involve data sets of one million records or more. Big Social Data are the high-volume, high-velocity, high-variety, high-value and high-variability data generated on social media (Alaoui et al., 2018) involving one million data points or more. Big Social Data are accessible through i.e. social monitors and listening tools such as SentiOne.

3.2.1.2. SentiOne

SentiOne is a social listening tool that allows its users to collect both historical Social Media Data and real-time Social Media Data (SentiOne, 2018a; Crunchbase, 2018). It is listening to and analyzing the Internet in 31 countries and 27 languages[16]. A prerequisite for understanding the study is to understand how SentiOne analyses and categorizes the historical and real-time data.

Through SentiOne's *Project* section, search queries can be set up (called '*keywords*'), which SentiOne uses to retrieve the mentions of interest from the Internet (SentiOne, 2018a). Users can choose between five different kinds of projects: *Brand*[17], *Social Profiles*[18], *Other*[19] and two different kinds of *Advanced Project Configurators*[20]. Depending on the choice of project type, different requirements will be set to run the project configuration. For instance, in order to set up one of the two types of the Advanced Project Configurator, the user will need to go through 4 steps and make specifications if needed in all of those steps. These are: *Add Keywords, Add Facebook, Add Sources, Set up additional Options*. If all steps are well specified and the project configurator runs the download file, which will be in a CSV or Excel format depending on the user's priority, the following variables are specified for *each* mention on social media that matches the requested keywords:

- *ID:* each post has a unique ID. Hence, if two identical publications are posted on i.e. Facebook *and* Twitter, the ID will be different for the two posts.
- In the column '*Link to source*', the URL to each of the relevant mentions is indicated.
- The *title* indicates the title of the publication. Not all publications have titles.
- *Content of post* includes the content that is posted. In the data extract from SentiOne there is a maximum of how many characters to be included. However, if the user wants to read the entire publication, it can most often be found by clicking on the *link to the source* (unless the URL has changed or the publication is removed)
- *Keywords* indicate which keywords the mention matches with. Often when a user is running Advanced Configuration Projects, more keywords are selected. In this column, the user can quickly get an overview of which publications that belongs to which keyword.

---

[16] Austria (de-at), Bulgaria (bg), Bosnia and Herzegovina (bs), Croatia (hr, sh), Czech Republic (cs), Denmark (da), English (en, en-gb, en-ie), Estonia, Finland (fi), French (fr, fr-be, fr-ch), Germany (de, de-at, de-ch), Greece (el), Hungary (hu), Italy (it, it-ch), Latvia (lv), Lithuania (lt), Montenegro (me), Norway (no), The Netherlands (nl, nl-be), Poland (pl), Portugal (pt-pt, pt), Romania (ro), Russia (ru), Serbia (sh, sr), Slovakia (sk), Slovenia (sl), Spain (es-es, es), Switzerland (ch) and Sweden (sv), Ukraine (uk) (SentiOne, 2018b).

[17] *"..this is is the easiest way to verify what people are saying about your brand, product or competitors online"* (SentiOne, 2018a)

[18] *if you add your social profiles (e.g. Facebook, Instagram, Twitter) you can have all your social mentions in one place"* (SentiOne, 2018.

[19] "*This type of search is a space for you to use your imagination and find whatever pops into your mind!"* (SentiOne, 2018a)

[20] *"Two types of search within the Advanced Project Configurator where you can use many advanced rules*" (SentiOne, 2018a)

- *Project name* defines the name of the project. It will be the same for all the publications, since it is in the same project.
- *Created* indicates the day the publication was created.
- '*Type*' can be either an *article* or a *post*: each relevant publication on social media is analyzed by SentiOne and grouped into either the category of being an 'article' or being a 'post' (SentiOne, 2018c).
- *Domain group* can be either *Facebook, Twitter, blogs, review, portal, forum* and *photo&video*. All articles and posts are grouped into a domain group depending on which media the content is published on. If an article or post is published on Facebook, SentiOne will group the article/post as 'Facebook' in the column. If instead the article or post is published on Instagram *or* Youtube, the article/post will be grouped as 'photo & video' in the domain group column.
- *Sentiment group* can be *positive, neutral* or *negative*. SentiOne groups the user's data extract into one of those three types of sentiments. The sentiment analysis is based on SentiOne's own developed algorithms, which are based on the PANAS schedule[21] developed by John R. Crawford and Julie D. Henry (2004). It is, however, important to mention that SentiOne's technology is not able to determine the sentiments of articles. Therefore, these fields are left blank in the column 'Sentiment group'.

3.2.2. *Store, structure and pre-process data*

The second step in J. P. Morgan's (2017) process is to *store, structure and pre-process data"*. The storing of big data often requires a huge amount of space to save the data. This can either be done on a hard desk or in the cloud. Once the data are stored, the structuring and pre-processing of the data require either high technical skills or advanced data preparation software. The data preparation and pre-processing include various cleaning, - and transformation activities, which is explained in the following four sections.

3.2.2.1. Data deduplication

[21] "*The PANAS Schedule is used to evaluate the frequency and intensity of the experience of positive and negative affective states, which constitute one of the basic components of happiness and subjective well-being. Negative people see the world in darker colors. NA and PA scales reflect aspects of the disposition of a subject. High NA points to the subjective anguish and unpleasant obligation, and low NA to lack of experience. In contrast, PA shows when a person experiences a pleasant engagement with the environment. Thus, emotions such as enthusiasm and alertness are indicators of high PA, while low PA characterizes lethargy and sadness*" (SentiOne, 2018a)

Many scholars in computing science state that data deduplication is crucial when working with any kind of data (Druva, 2009). Data deduplication is a data compression technique for removing duplicate copies of repeating data (Druva, 2009). Hence, duplicates must be identified and deleted. Some data preparation software (e.g. Alteryx) has built-in tools that are capable of identifying and excluding duplicates.

3.2.2.2. Filtering

Filtering is a data cleaning approach in which data are scanned for relevant and irrelevant information (Munzner, 2009). Social Media Data extracted from social monitors can be corrupted by *noise* and *outliers* occurring from errors in the identification or acquisition of data (Garcia, Carvalho and Lorena, 2013). Typically, filtering reduces noise, whereas outliers usually are detected through *visualization techniques* (see more in subsection 3.2.2.4.). An outlier is "*..an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*" (Barnett and Lewis, 1978). Consequently, filtering out irrelevant data is a crucial element in the pre-processing. If that part is not done thoroughly, it may have fatal consequences for the results, and eventually the conclusion and recommendations.

3.2.2.3. Time adjustment

Time adjustments can be required in the data pre-processing to provide coherent and consistent data. Time adjustments may include a transformation i.e. the whole raw data set reported on a daily data, a weekly, a monthly, a quarterly or a yearly basis. When there is an absence of data in some time periods, it can impact the analysis. Therefore, it is important to make the right time adjustments. In addition, the data should be skimmed for *seasonality* and be adjusted accordingly. Seasonality is a characteristic of a time series where the data experiences predictable change on a regular basis (i.e. each year). To adjust the data, each data point should simply be divided by the seasonal factor of its month/quarter/year.

3.2.2.4. Visualization

According to Thomas and Cook (2005) "*Visual representations translate data into a visible form that highlights important features, including commonalities and anomalies*"(p.69). Consequently, visual data representation allows researchers to spot trends or outliers easily that would be impossible to spot in the raw data set. It improves the understanding and insights into complex datasets in complex situations where human visual capabilities are needed (Munzner, 2009). The goal of visual analytics is to picture

natural phenomena in an accessible and appealing way (Heer and Segel, 2010). Visual representations are accessible through visualization analytics software such as Tableau and Excel.

3.2.3. *Analyze data*

The third step in J. P. Morgan's (2017) process is *"Analyze data"*. A wide range of data analysis methods and techniques are described in the literature. However, in relation to this study, it is relevant to understand what *sentiment analysis* is, to explain the *regression analysis* and *p-value and hypothesis testing*.

3.2.3.1. Sentiment analysis

Sentiment analysis is the process in which feelings, moods or opinions in a text bid are computationally identified and categorized (Pang, 2008). Sentiment analysis is a suitable approach to determine people's attitude towards products and companies (Choi and Varian, 2012) or politicians (Tumasjan et al. 2010). Sentiments can be *positive*, *neutral* or *negative* (Corea and Cervellati, 2015) and is often processed through NPL. The polarity of a message is determined in different ways. Some software is programmed to look exclusively into the text, whereas other software is capable of analyzing the emoticons added to the message. Some computers manage to categorize the polarity based on both the text and the emoticon. Emoticons are representations of a facial expression such as '🙂' (expressing a smile) or '❤️' (expressing love). The constant improvement in the automated computational sentiment classification has increased the quality of sentiment classifiers recently. As of today some social listening tools even provide real time sentiment classifications of messages, which is used by the present study, and thus, an in-depth explanation of the technical aspect of sentiment analysis is not provided in the study.

3.2.3.2. Regression Analysis

Regression analysis is the statistical process of investigating relationships between a *dependent variable* and one or more *independent variables* (Darlington and Hayes, 20, p.1). It estimates the unknown effect of one variable over another (Stock and Watson, 2003).

There are three types of regression models: *linear regression models*, *non-linear regression models* and *polynomial regression models* (Pellakuri et al., 2015). Similar to Bissattini and Christodoulou (2013), Ruiz

et al. (2012) and Lassen, Madsen and Vatrapu (2014), the present study uses linear regression analysis. There are two types of linear regression models; first, the simplest form is the *simple linear regression* (Darlington and Hayes, 2017). If *one* independent variable is chosen, it is a simple linear regression model. Second, the *multiple linear regression model* (Darlington and Hayes, 2017) allows a single variable to be predicted from a *set* of independent variables. The simple linear regression formula is defined as:

$$f(x) = bX + a$$

Whereas, *f(x)* is the estimated value of the dependent variable, and *b* is the slope, also called the regression coefficient and *a* is a constant value of the dependent variable when the independent variable *x* is zero.

The general multiple regression formula in correlation analysis is defined by the formula:

$$f(x) = a + b_0x_0 + b_1x_1 + b_2x_2 + \ldots b_nx_n$$

Where *f(x)* is the estimated value of the dependent variable, and b_0 is the regression coefficient for independent variable x_0, b_1 is the regression coefficient for independent variable x_1, and so forth. *a* is the Y-intercept.

3.2.3.3. P-value and hypothesis testing

In order to analyze the output of a linear regression model, R. A. Fisher[22] suggests that the variables in the statistical model should be tested for *significance*. To understand significance, it is a prerequisite to understand the concepts of hypothesis testing. In statistics, there are two types of hypotheses: *the null hypothesis* and *the alternative hypothesis*. The null hypothesis states that there is no significant difference between the observations, and if a significant difference exists, it is due to sampling or experimental errors (Banerjee et al., 2009). The alternative hypothesis states that there is a statistically significant relationship between two or more variables (Banerjee et al., 2009). To determine statistical significance in a hypothesis test, the confidence level must be decided. The 95% confidence level is mostly used in research papers (Zar, 1984), and reflects a significance level of 0.05 (Field, 2013). If a true parameter value is X but the 95% confidence interval does not contain X, then the estimate is significantly different from X at the 5% significant level, and the null

[22] Sir R. A. Fisher was a British statistician and is perceived to be the father of the modern statistics.

hypothesis can thus be rejected. After the confidence level is decided, the F-statistics or test statistics can be used to generate p-values*: "..the p-value represents the likelihood of getting our test statistic or any test statistic more extreme, if in fact the null hypothesis is true"* (Keiser Education, 2017). If the null hypothesis does not provide a plausible explanation of the data (based on the p-values generated in the F or test statistics), there is statistically significance for rejecting the null hypothesis. Some common decision rules in the literature concerning the p-values exist:

- ≤ 0.05 scholars declare their data to be *significant,* and the null hypothesis can be rejected;
- > 0.05 scholars declare their data to be *not significant,* and they fail to reject the null hypothesis.

However, in hypothesis testing one must be aware of *type I errors* and *type II errors.* Type I error is the event of rejecting the null hypothesis when it is true (Banerjee et al. 2009), and the Type II error is the event of not rejecting the null hypothesis when it is false (Banerjee et al., 2009). The probability of getting a type I error is 5% (for confidence interval = 95%) and the probability of getting a type II error is also 5% (Cohen, 1998). It is due to the threshold at 0.05, which means that the study accepts 5% probability of identifying an effect when there is not one. The chance of getting a type II error decreases with the number of tests performed, whereas the chance of getting a type I error increases with the number of tests performed. Bland (2000) emphasizes the importance of testing for type I errors as the probability of getting insignificant null hypothesis increases with the number of tests performed. If a test uses the significance level of 0.05 and includes two independent true null hypotheses, the probability of neither of the tests will come out significantly is 0.95 times 0.95 = 0.90 (Bland, 2000, section 6.2). Testing twenty of such independent true null hypothesis, the probability that none will be significant is $0.95^{20} = 0.36$. Hence, the probability of getting at least one significant result is: 1-0.38 = 0.64, and thus, the likelihood of getting one significant result is higher than not getting one.

3.2.4. *Trade Ideas*

The *Trade Ideas* part is the execution of the trades. In the present study, the step *trade ideas* is limited to the scope of the research question of the project and thus focuses solely on understanding investors' degree of use of social media trading and the relationship between social media content and stock market performance indicators. Hence, it is not elaborated in this study whether or not investors should trade the ideas presented in this study.

4. Theoretical framework

A theoretical framework "*introduces and describes the theory that is used to examine the research topic*" (Gibson, 2016, p. 2). Theoretical frameworks are useful to frame the perspective of the study and will help researchers to investigate the chosen objective of the analysis (Stokes, 2013, p. 64). This study has chosen the AIDA model as the overall theoretical framework for structuring and analyzing investors' familiarity with social media trading. The AIDA model is also applied to analyze the Big Social Media Data and the Stock Performance Data.

4.1. *AIDA framework*

The new source of information created by big social data creates the opportunity for investors to produce meaningful facts, actionable insights and ultimately economic outcome for them. Using the AIDA framework, it can be deducted that attention may lead to purchases (Hassan, Nadzim and Shiratuddin, 2015). This is particularly interesting for the banking sector and the entire investment industry, which are characterized by conservatism and a low willingness to adopt new trading methods due to risks (Berger and Woitek, 2001). While big social data in interplay with Stock Performance Data can be used in multitudes of analyses, this study solely focuses on the investigation of a possible correlation between the attention, interest and desire a specific Danish listed company achieves on social media and the purchases (action) of the company's stock on the stock exchange. Therefore, this is a study of real-world events. This study's application of AIDA aims to investigate the degree of investors' usage of social media trading and if their attention on social media trading leads to purchase decisions of stocks. Influential factors that may lead a person from one phase in the AIDA model into another is explained below.

Entering the first phase: *awareness/attention* of social media trading or the stock or the product or service the listed firm offers can be a consequence of:

- Reading news either online, on social media or offline;
- Learning that friends, family or colleagues use it, have it or have preferences towards it;
- Watching commercials;
- Seeing it in use;
- Observing the performance of it;

The individual can now proceed to the second phase. Entering the second phase, *Interest/knowledge* of social media trading can be a consequence of:

- Searching for more information either on social media, online or offline;
- Reading reviews on social media, online or offline;
- Seeing role models using it;
- Hearing role models suggesting to use it;
- Seeing the results and the effects of using it;
- Comparing the results of using it to alternatives;

The individual can now proceed to the third phase. Entering the third phase: *desire/preference* of can be a consequence of:

- Hearing other persons' experiences with it and on that basis forming preferences towards it;
- Evaluating how it can strengthen or improve the job or life;
- Deciding on whether it is superfluous, nice to have or need to have;

The individual can now proceed to the fourth phase. Entering the fourth phase, *action/purchase* can be a consequence of:

- High preferences towards it, which create the needs to purchase it;
- Low preferences towards it, which create the need *not* to purchase it;
- Low preferences towards it which create the need to act on another trading strategy, stock, product or service offered by another firm;

The AIDA model is illustrated in Figure **2**.

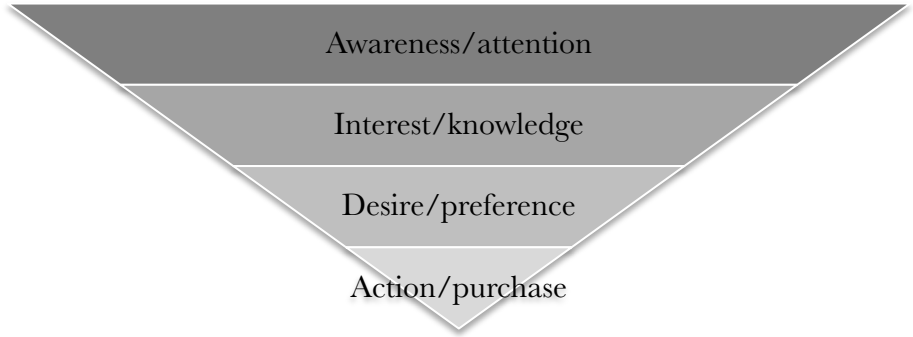

Figure 2: AIDA model

5. Methodology

"*Research methodology is a systematic way to solve a problem (…)and describes how research is to be carried out (…) Its aim is to give the work plan of research"* (Rajasekar, Philominathan, Chinnathambi, 2013, p. 5). Furthermore, research methodology considers suitable methods to solve problems and efficiencies of the chosen methods (Rajasekar, Philominathan, Chinnathambi, 2013).

For this study, each of the steps in the AIDA model serves as a guideline to the creation of the methodological tools. This study both uses qualitative data and quantitative data. Qualitative data that cannot be quantified, and thus, they are used to explain and understand phenomena (Veal, 2001). It often originates from interviews or focus groups (University of Arkansas, 2018). Quantitative data are numerical or statistical data that often origin from surveys, experiments, instruments, tests or questionnaires (University of Arkansas, 2018).

To answer research question 1 and research question 2, investors' awareness of, interest in, desire for and action on social media trading are investigated by applying both qualitative and quantitative data. The use of triangulation[23] increases the credibility and validity of the results as it cross-checks data from multiple sources (O'Donoghue and Punch, 2003), and thus, strengthen the results.

Due to large size stock trades executed by professional investors, and hereby greater power of market movements, this study finds a qualitative approach most efficient as it provides greater detail and depth of information about individuals' experience and understanding of events, which is necessary to understand in order to find out in which phase of the AIDA model the professional investors are (Kvale, 1996), and through this understanding answer research question 1. By conducting interviews with the professional investors, this approach allows the researcher to deeply understand their attention to, interest in, desire for and action on social media trading and which factors that may prompt their entrance into a new phase in the AIDA model.

Due to smaller stock trades executed by each individual private investor, and hence, limited power to push stock markets, this study finds a quantitative approach most efficient for understanding in which phase of the AIDA model the private investors have entered, and through this understanding answer research question 2. By creating a questionnaire for the private investors, this approach allows the researcher to understand the *mass'* attention to, interest in, desire for and action on social

[23] Triangulation means using more than one method to collect data on the same topic (ResearchGate, 2014).

media trading. The in-depth understanding of all the respondents would not have been possible due to time constraints.

In order to answer research question 3 and research question 4, the study uses Big Social Data retrieved from SentiOne and Stock Performance Data retrieved from DataStream. Both sources deliver the data in a quantitative format. The collection of Big Social Data would have been impossible with a qualitative approach.

The present study is mainly concerned with the generation of new theory emerging from the data collected, and one can, thus, say that the overall approach of the study is inductive (Veal, 2011).

The data used for this study include both primary data and secondary data. Primary data are data that are collected by the researcher himself/herself for a specific purpose (Veal, 2001). The data collected firsthand by the researcher of the present study include interviews with professional investors and questionnaires for the private investors.

Secondary data are data, which are collected by someone else for some other purpose (Veal, 2001). This study uses secondary data from SentiOne to collect Social Media Data and from DataStream to collect Stock Performance Data. The usage of both primary and secondary data adds credibility to the study (Veal, 2001; O'Donoghue and Punch, 2003).

5.1. *Primary data*

5.1.1. *Qualitative Interviews with professional investors*

The qualitative interviews are designed and prepared to interview Danish professional investors entitled as *investment officers* or *chief strategists* in large or medium size Danish investment firms.

All investors that qualified to be interviewed are contacted via e-mail. A brief presentation of the researcher, the definition of social media trading and the scope of the study are given in that first email (see appendix 1). The subject title of the e-mail is 'CBS student with few questions for the Master Thesis'. All interviews are conducted by telephone interviews in the period: March 14th to April 4th 2018. All interviews are sound recorded and transcribed manually in Microsoft Word with the acceptance of the interviewee. After the transcription, the interview is sent to the interviewee who accepted the content of the interview in order to avoid misunderstandings that possibly can influence the outcome of the study.

5.1.1.1. Design of the interviews

The design of the interviews is constructed to answer each of the four parts in the AIDA model namely the *awareness* of the possibility of using social media for stock picking, the *interest* in using social media for stock picking, the *desire* to use social media for stock picking and the actual *action* of using social media trading for stock purchase. The interviews are designed as semi-structured interviews (Veal, 2001) aiming to answer the first research question. The interview protocol includes 10 open-ended questions that help the interviewer to guide the interview (see interview guide, appendix 2). This interview design maintains some structure, and works more as a guided two-way conversation between the interviewer and the interviewee. However, it also enables the interviewer to probe the interviewee for additional details, and the interviewees are allowed to express themselves in their own terms, as the semi-structured interview offers a great deal of flexibility for the interviewer and at the same time making sure that the interviewer will keep focused on gathering all the information that is needed to answer the research question. New questions are asked during the interview and some of the following pre-arranged questions can be excluded if they are found to be superfluous. The first question asks if the interviewer may record the interview. The second question asks directly if the investment firm uses social media trading to decide on investments. The next two questions focus on the company's awareness of the possibility of using social media trading in the financial sector. Question five and six focus on the company's interest in social media trading for investment and trading purposes. Question seven and eight focus on which of the company's needs that could be satisfied or have been satisfied by the usage of social media trading in the firm. Question nine and teen focus on the company's action on social media trading when making investment decisions. Once the interview guide protocol is completed, one interview is pilot tested and changes are made accordingly.

5.1.1.2. Sampling

A *selective sampling method* is appropriate for the conduction of the interviews as the researcher aims to select experts with a certain degree of knowledge of their employers' current and future trading strategies. Qualified persons for the interview are persons entitled as 'chief strategists' or 'investment officers' within Danish large or mid-size investment firms in Denmark. Screenings on LinkedIn and Google searches are used to identify qualified persons.

Since the list of qualified persons is rather long, it is necessary to narrow down the number of people to be contacted by setting up some basic characteristics for the person to qualify for the interview. The characteristics are: the employee should have been with the firm for more than one year; the person should speak Danish and should be willing to attend a telephone interview. One person from

each company was randomly selected and contacted. In some cases this person forwarded the invitation to another person in the organization. As the person that fulfilled the requirements recommended the other person, the new person did not have to meet the requirements stated above. Out of the list of 21 firms, 20 firms were contacted per e-mail and asked to participate in the study. Four out of the 20 firms showed willingness to participate and they were all interviewed. These companies were: Nordea, Danske Bank, Carnegie and Pension Danmark. The characteristic summary of the respondents is presented in Figure **3**. The researcher knows the full names of the interviewees.

Company name	Name	Position
Danske Bank	AV	Strategist
Nordea	PJ	Chief Strategist
Carnegie	HD	Chief Strategist
Pension Danmark	FV	Chief Strategist

Figure 3: Characteristic summary of the professional investors

5.1.2. *Questionnaire to private investors*

To complete the analysis of the investors, and to answer research question number 2, the study gathers quantitative data from an online questionnaire. The method of data collection is considered as appropriate because it is easy to reach many investors without spending much time or money, and at the same time getting a relatively high number of answers.

5.1.2.1. Design of the questionnaire

The aim of the questionnaire is to investigate to what extent social media influence private investors' decision-making process.

The questionnaire is entitled as "Private investors' use of social media trading for trade decision making" (see appendix 3), and contains initially two closed questions, which is followed by respondent-completion questions regarding the private investors' use of social media trading strategies. In order to be able to submit the survey, all fields have to be answered. It is only possible for the respondents to submit one answer per IP address. SurveyMonkey is used as the survey deliverer and a short message is given with the link in order to increase the probability that the private investors would start and complete the survey. The message with the link was: 'Hi everyone, I am a student writing my Master Thesis about investors' usage of social media when doing their

stock picking. I would be very grateful if you would take a few minutes to answer the survey anonymously. It can be found in the link below. Thank you so much". Once the questionnaire is completed, it is pilot tested with three private investors from the researcher's own network and changes are made accordingly.

5.1.2.2. Sampling

The message with the link is posted on Scandinavia's largest investor network on Facebook, Aktieporteføljen on 3 April 2018 and is accessible until 8 April 2018. The total number of members of Aktieporteføljen counts more than 44 thousands. The number of answers was 98.

5.1.3. *Interview with statistics expert*

To validate the methodology of the secondary data analysis and to interpret the results correctly, an interview with a statistics expert is conducted. As the conclusion to research question 3 and research question 4 depends on the regression analysis results and its interpretation, this interview is initiated to strengthen the validity of the findings of the study. The interview design works as an open discussion between the interviewer and the interviewee. The interview is a Skype meeting and is conducted on 13 April 2018. Four days before the interview takes place, the interviewer sends the regression model results to the expert to make him prepared for the interview. The interviewee is aware of his contribution as an expert in the thesis (see appendix 4). No recordings are made, but notes are taken throughout the interview (see appendix 5).

5.1.3.1. Design of the interview

The interview is designed as a semi-structured interview (Veal, 2011), and the interviewer designs the interview protocol prior to the interview. The interview protocol includes 12 open-ended questions that help the researcher guide the interview (see interview guide, appendix 6). The open-ended interview design maintains some structure, and secures that all relevant results are evaluated. This design enables the researcher to probe the interviewee for additionally details, and likewise, and the interviewee is allowed to express himself in his own terms, as the semi-structured interview offers a great deal of flexibility for the interviewer. At the same time this design makes sure that the interviewer keeps focused on gathering all the information that is needed to validate the results. New questions are posed during the interview and some of the following pre-arranged questions are not asked as they turn out to be superfluous.

5.1.3.2. Sampling

The selective sampling method is chosen for the data collection as only experts with a special knowledge of regression analysis can contribute to the study. The researcher of this study screens all her previous teachers' CVs at CBS to learn about their competences within regression analysis and statistics. One person, Bersant Hobdari, Associate Professor in Department of International Economics and Management, has a CV with the relevant qualifications within regression analysis. The researcher contacted Mr. Hobdari by e-mail (see appendix 7) and arranged an interview.

5.2. *Secondary data*

Two secondary data sources are used in this study. Those are: *Big Social Media Data* and *Stock Performance Data.*

5.2.1. *Big Social Media Data and Stock Performance Data*

Social Media Data are collected in order to see if social media users are *aware* of, *interested* in or have a *desire* for the products, services or stocks of Danish listed companies. If there are social media mentions about a firm, the study concludes that social media users have entered all three phases in the AIDA model. It is not evaluated if the *same* social media users also purchase the firm's stock. The stock performance indicators are used to investigate if any *activity* on the stock purchase takes place. It is then the aim of this study to investigate if any relationship between the social media mentions and the stock performance exists. To do that, the Social Media Data and stock performance indicator data processing follows J. P. Morgan's (2017) stepwise approach explained in the section '*Types of data, data collection concepts, data analytics techniques and models*'. This section is divided as illustrated in Figure **4**.

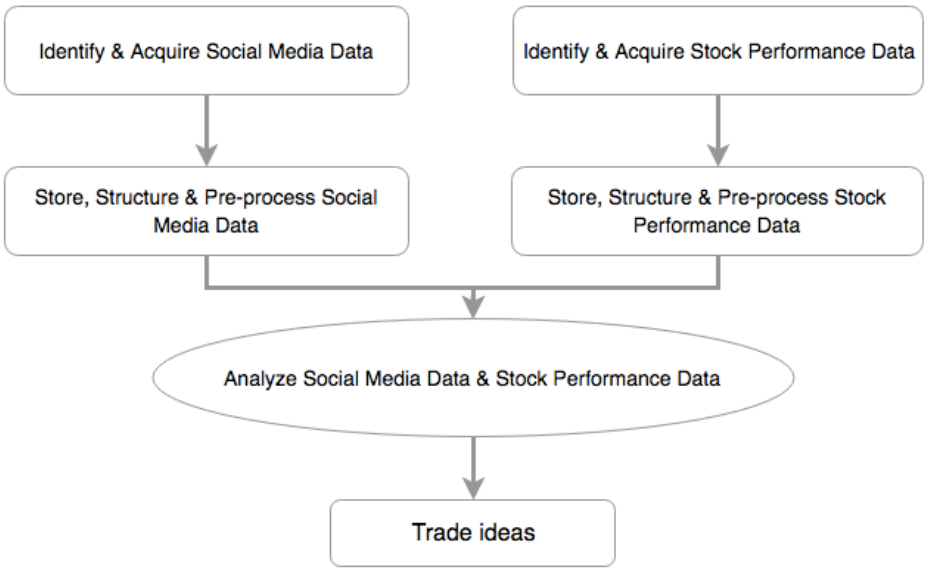

Figure 4: The thesis' application of J. P. Morgan's (2017) stepwise approach

5.2.2. *Identify and Acquire Big Social Media Data*

This paper identifies the Danish C20 firms as of 19 January 2018[24] (Børsen, 2018) to be studied. To acquire the Social Media Data, the most common social media mentions about the C20 firms are identified (see appendix 8) through intense searches on all the social media SentiOne supports. Once the mentions are identified, the data collection can start.

This study uses SentiOne for the data collection. The common mentions on social media about each of the C20 firms are entered in the SentiOne Project Configurator, which means that when the SentiOne Project Configurator connects to the SentiOne database, it will look for only the mentions specified. Before the SentiOne Project Configurator is set to start running the data download, the researcher specifies a few more details: the data collection period ranges from 31 December 2014 to 19 January 2018. The collection includes social data from major domain groups including Instagram and Youtube (Instagram and Youtube are from now on called 'photo & video), blogs, portals, Facebook, reviews and Twitter. The data collection includes all mentions from all languages supported by SentiOne. The data collection is set to include the sentiments of the post, which can be positive, neutral or negative. The raw Social Media Data was downloaded as a CSV file. The steps are visualized in the process flow diagram in Figure **5**.

[24] A. P. Møller - Maersk, Bavarian Nordic, Carlsberg, Chr. Hansen Holding, Coloplast, Danske Bank, DSV, FLSmidth, Genmab, GN Store Nord, ISS, Jyske Bank, Lundbeck, Nordea Bank, Novozymes B, Pandora, Vestas Wind Systems, Ørsted

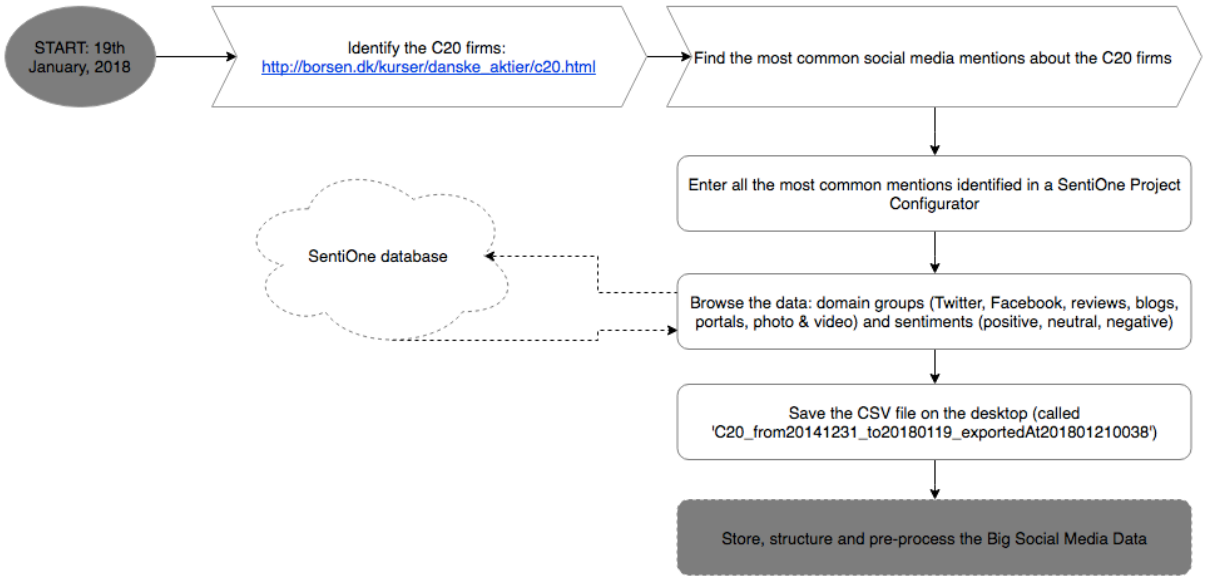

Figure 5: Process flow diagram: identify and acquire the Big Social Media Data

The last step in the process *Store, structure and pre-process the Big Social Media Data* is explained in detail in the following section.

5.2.3. *Store, Structure and Pre-Process the Big Social Media Data*

To limit difficulties in reading the CSV file in the data analytics software Alteryx, the CSV file with the raw data is imported into Python by using Sublime Text. In here, suitable delimitators are set up in order to structure the data. The Python inputs can be found in appendix 9. Furthermore, the CSV includes a large range of data, but the following columns are extracted and selected for further analysis: *ID, title, link to source, domain groups, type, content of posts, keywords, sentiment, project name,* and *created*. The file is hereafter saved as an Alteryx file.

The data cleaning is done by applying Alteryx. The Alteryx file described above is the input data, which contains all the Social Media Data. The data are filtered in Alteryx. The first filter applied is the *unique: ID*, meaning that the same mention only can appear a single time on the same social media. Hence, the same mention can occur more times if it is posted on more social media. Hereafter, the *select* filter is chosen to rephrase '*created*' to only indicate the *date* and not the *date and time*. Hereafter, the *formula* filter is used noting the *output column* as '*clean_keyword*' and filtering on the *keywords*. In contrast to *keywords*, *clean_keyword* can only contain a single keyword. If a keyword contains both "Genmab" and "danskebank", it structured by using the filtering formula: *if Contains([Keywords], "genmab") THEN "genmab" ELSEIF Contains([Keywords], "danskebank") THEN*

"danskebank" and so forth. For each *clean_keyword* that includes one of the company names, an Alteryx file is created. The new files generated are further opened and cleaned separately in Alteryx. Some of the variables including *author, keywords, sentiment points, tag, gender, project name* and *added to system* are removed as their correlation with the stock prices is not in the scope of this paper. The cleaning process is initiated with a subjective discretion over noising elements in the *content of posts* column and the titles of the articles (i.e. in the dataset of Pandora, it appeared that i.e. Disney World in Orlando has a theme park that is called 'Pandora - The World of Avatar'; a gaming and music streaming platform that is called 'Pandora Gaming'; in the reality program Paradise Hotel, which runs in various countries around the world, 'Pandora's Box' is a weekly event). Please see appendix 10 to learn more about which noising elements that are filtered out for each firm. As all languages supported by SentiOne are accepted, the subjective discretion leads in some cases to limited understanding of the content of the post or the title of the article. In those cases, the subjective discretion is made on the *link to source*. If the link to the source does not work or is irrelevant for the firm (i.e. the dataset of ISS, many sources linked information about the ISS NASA, which is completely irrelevant for the Danish listed firm ISS), the mention is filtered out.

Hereafter, the data cleaning is performed by Alteryx. This process of subjective discretion and data cleaning process in Alteryx continues until no noise occurs in data samples of 200. If noise repeatedly emerges in the dataset, and no pattern of the noising elements is detected, the dataset is evaluated as being too risky to include in the analysis. The variables *title, link to source, domain groups* in the clean datasets are hereafter removed, and the files are saved in an Excel format (see appendix 12). Hereafter five firms are randomly chosen to be studied further.

When the selected data are cleaned properly, the social data transformation begins. It requires time adjustments and transformation of variables to percentages.

This study uses the *financial calendar* year as referred to in Lassen, Madsen and Vatrapu (2014) as the time format. On that basis, the quarters in a year are defined as: first quarter (Q1) includes all days from the 1 January to 31 March. Second quarter (Q2) is from 1 April to 30 June. Third quarter (Q3) runs from 1 July to 30 September, and fourth quarter (Q4) runs from 1 October to 31 December. Initially, the Social Media Data are reported on a daily basis. It is therefore required to make time adjustments to the Social Media Data variable *created* to transform it into a quarterly character. Time adjustments are done by using the '*VLOOKUP*' function in Excel. The first part of the Big Social Media Data pre-processing is hereafter done.

For each of the category variables: *type, domain group* and *sentiment*, the different types of content, domains and sentiments are counted. In the data downloaded from SentiOne, the *types* can either be *article* or *post.* A *domain* can either be: *Facebook, Twitter, Blogs, Forums, Reviews, Portals* or *Photo and Video.* A sentiment can either be *positive, negative* or *neutral.* For each category variable, the total number of each variable in the specific time periods is counted by using the *COUNTIFS* function in Excel. This function enables the researcher to count the sum of e.g. different kinds of domains in a column range that meets the criteria of being created in e.g. the second quarter in 2017, being positive and posted on Facebook. When this is done, the three most used domains are identified, and the amounts of positive and negative sentiments on each domain for each quarter are counted. These three will as well as the *positive sentiments, negative sentiments, articles* and *posts* later be used as the independent variables in a simple linear regression analysis. However, before this can happen, these variables must be pre-processed further. The social media variables are transformed into the change in the sum from one quarter, *t* (i.e. Q216) to next quarter, *t+1* (i.e. Q316):

$$\Delta\sum positive\ sentiments_{t+1} = \frac{\sum(positive\ sentiments_{t+1}) - \sum(positive\ sentiments_t)}{\sum(positive\ sentiments_t)}$$

The numbers of variables that are pre-processed as described above adds up to a total of 19. These variables are: *change in number of articles, change in number of posts, change in number of positive sentiments, change in number of negative sentiments, change in number of blogs, change in number of forums, change in number of portals, change in number of tweets, change in number of facebook, change in number of reviews, change in number of photovideo, change in number of positive posts, change in number of negative posts, change in number of positive portals, change in number of negative portals, change in number of positive tweets, change in number of negative tweets, change in number of positive facebook, change in number of negative facebook.*

When the numerical pre-processing is completed the new variables as functions of time are translated into a visible form as Thomas and Cook (2005) suggest. The visual representation is done in Microsoft Excel. Rousseeuw an Hubert (2011) advises researchers to be careful about excluding outliers. However, if the outlier is erroneous, then the outlying value should be deleted from the dataset.

The storing, structuring and pre-processing of the Big Social Media data is now done and the analysis of Social Media Data and the Stock Performance Data can begin. The flow of the *store, structure and pre-processing of the Big Social Media Data* is visualized in Figure **6**.

The last step in the process *Analyze data* is explained in detail in the subsection '*Analyze data'.*

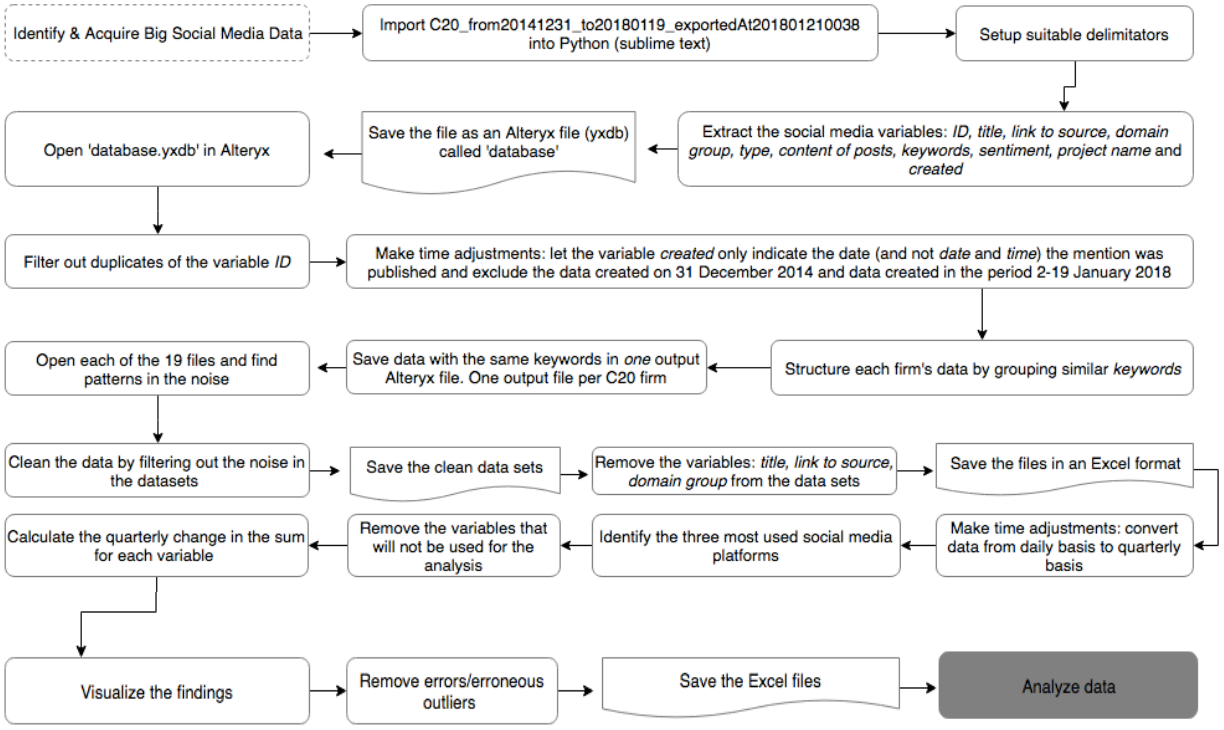

Figure 6: Progress flow diagram: store, structure and pre-process the Big Social Media Data

5.2.4. *Identify and Acquire Stock Performance Data*

The historical Danish stock data are collected through Datastream. In the *Datastream Request page*, the selections are set and the searches are run. The Datastream Navigator provides two types of Navigators that are used for the data collection. Firstly, the *navigator* is used to find the bundles of the Danish stocks of interest. One option in the drop down menu is chosen: "NASDAQ C20". Secondly, in the *Datatype Navigator*, four specific stock market indicators similar to the ones used by Park and Irwin (2007), Wei, Chen and Ho (2011), Boykin (2017), Rahman (2011), and Shamsudin, Mahmood & Ismail (2013) are chosen, namely: opening price, volume traded, price/earnings ratio and price/book value. In the *Settings* window, the *timeframe* is set to start on Wednesday 5 February 2014 and end on Wednesday 7 February 2018. The *frequency* is on a weekly basis, meaning that the Stock Performance Data items for each stock are indicated on a week basis between the start and end date. After having specified the stocks, the Stock Performance Data and the time frame, the '*Run*' bottom is pressed. The output results are downloaded as an Excel file (see appendix 12). The process flow diagram is visualized in Figure **7**.

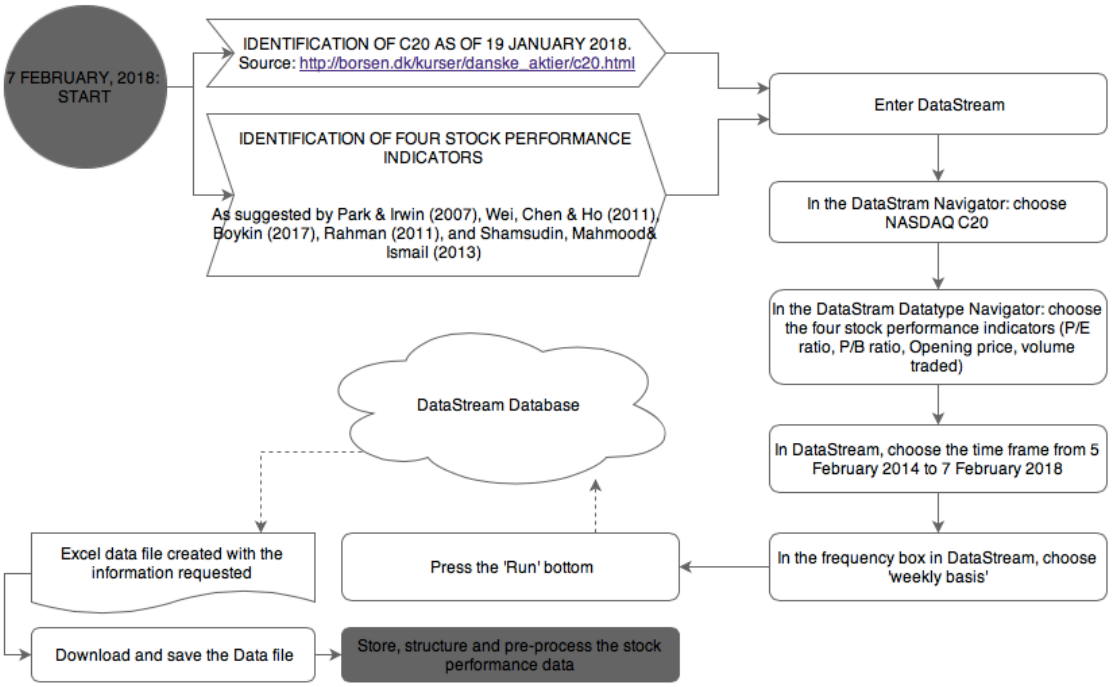

Figure 7: Process flow diagram: identify and acquire Stock Performance Data

5.2.5. *Store, Structure and Pre-Process Stock Performance Data*

Due to the well-structured data output of Datastream, the structuring and pre-processing of the data are quite easy and only consist of few parts. These parts relate to the time adjustment and the format of the stock performance indicators (opening price, volume traded, price/earnings ratio and price/book value). The time adjustment is done in two steps. Firstly, to copy each firm's data from the output data sheet that contains all of the data points and pasting it into a new Excel document: one Excel document per firm. Secondly, make time adjustments by transforming the data from *weekly basis* into *quarterly basis*. Previous studies including Lassen, Madsen and Vatrapu (2014) and Bagnoli et al. (1999) use quarterly data too. Before the practical transformation can begin, the starting date and ending date have to be defined. The definitions of the quarters are equal to that of the social media variables. As the frame for the transformation is now set, the transformation can be done by applying the '*VLOOKUP*' function in Microsoft Excel.

The pre-processing of the stock performance indicators is hereafter done. The quarterly reported data are transformed as follows: The *opening price* the first day in the quarter i.e. Q216 (called '*opening price*$_t$') and the opening price the first day in the next quarter, i.e. Q316 (called '*opening price*$_{t+1}$') is kept. The study uses the change in decimals. The change in opening price from quarter *t* to quarter *t+1* is calculated with the following formula:

$$\Delta opening\ price\ decimal_{t+1} = \frac{opening\ price_{t+1} - opening\ price_t}{opening\ price_t}$$

The same is done for the price to book value and the price equity ratio but the volumes of stocks traded are calculated differently. That is:

$$\Delta volume\ of\ stocks\ traded_{t+1} = \frac{\sum(volume\ of\ stocks\ traded_{t+1}) - \sum(volume\ of\ stocks\ traded_t)}{\sum(volume\ of\ stocks\ traded_t)}$$

In order to answer research question 4, the change in the stock performance indicators in the quarter following the change in the social media activity are noted as *t+2*, and calculated with the formula, i.e.:

$$\Delta opening\ price\ decimal_{t+2} = \frac{opening\ price_{t+2} - opening\ price_{t+1}}{opening\ price_{t+1}} \rightarrow \frac{opening\ price_{t+2} - \left(\frac{opening\ price_{t+1} - opening\ price_t}{opening\ price_t}\right)}{\left(\frac{opening\ price_{t+1} - opening\ price_t}{opening\ price_t}\right)}$$

The same method as before is used to calculate the change in the remaining three stock performance indicators.

When the numerical pre-processing is complete the new variables as functions of time are translated into a visible form, as Thomas and Cook (2005) suggest: if any erroneous outliers occur, these are removed from the dataset. The visual representation is done in Microsoft Excel. Hereafter, the dataset is ready for the next step: *Analyze data*, which is pictured in Figure **8**.

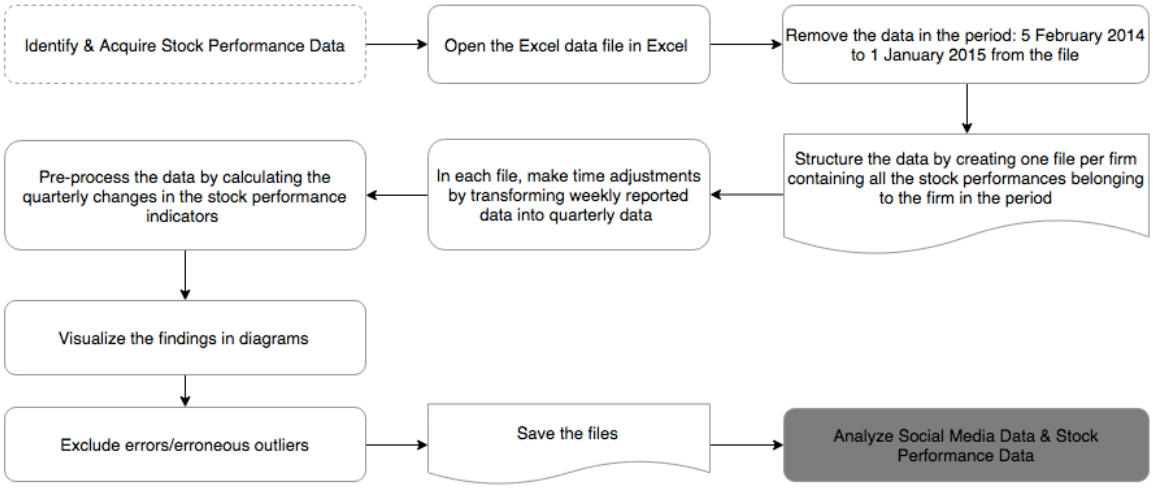

Figure 8: Process flow diagram: store, structure and pre-process Stock Performance Data

5.2.6. *Analyze Social Media Data and Stock Performance Data*

This part of the study aims to investigate if social media users' awareness, interest and desire have an impact on the stock purchases. It is done by regressing the social media variables described previously on the stock performance indicators described in the previous section.

To answer research question 3, social media variables and stock performance variables are gathered in the same period. In total, the impact of 19 social media variables on four dependent stock performance variables (*'change in opening price$_{t+1}$', 'change in volume traded$_{t+1}$', 'change in P/E $_{t+1}$', 'change in P/B $_{t+1}$'*) is investigated. This is done in order to measure if mentions on social media influence short-term stock performance. Likewise Lassen, Madsen and Vatrapu (2014), this study both performs simple, - and multiple linear regressions to answer research question 3. In total, 19 simple linear regression analyses are performed *per* dependent variable *per* firm, and 15 multiple linear regression analyses is performed *per* dependent variable *per* firm. The combination of independent variables for the multiple regression analysis is randomly picked (see appendix 11). It is visualized in Figure **9**.

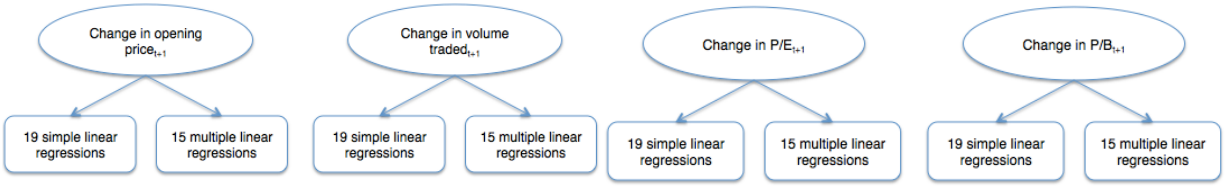

Figure 9: Regression analyses performed to answer Research Question 3

To answer research question 4, both simple and multiple linear regression analyses are performed as well. Social media variables from period *t+1* are gathered and stock performance variables from *t+2* are gathered to see if social media mentions in one quarter have influence long-term stock performance .In total, the impact of the 19 social media variables on four dependent stock performance variables (*"change in opening price$_{t+2}$", "change in volume traded$_{t+2}$", "change in P/E $_{t+2}$", "change in P/B $_{t+2}$"*) are investigated. Both simple, - and multiple linear regressions will be also performed to answer research question 4. From here, the same procedure is used as for research question 3. Please see Figure **10**.

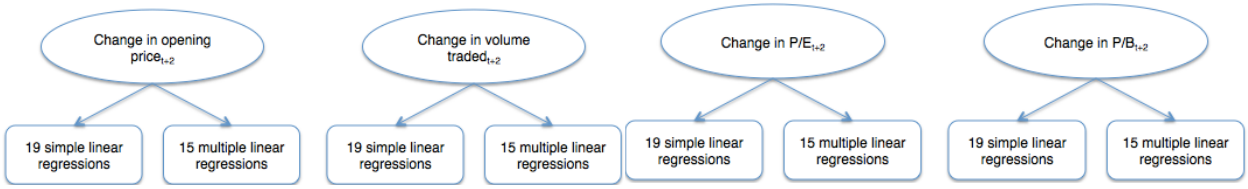

Figure 10: Regression analyses performed to answer Research Question 4

In consequence, this study will perform 76 simple linear regression analyses per firm (#independent variable times #dependent variables = 4*19 = 76) and 60 multiple linear regression analyses per firm (4*15 = 60) in order to answer research question 3. The number of regression analyses is the same for research question 4.

All regression analyses are done through the application of the Data Analysis and Statistical Software, STATA by applying the command '*reg*'. This is the first step in the analysis process. The *reg* output includes an *ANOVA table*, an *Overall Model Fit* and the *Parameter Estimates*. As suggested by Stock and Watson (2003, chapter 4) few things in the model are being checked before moving on to the next step including: *curvilinearity*[25], *heteroskedacity*[26] and *multicollinearity*[27].

It is in particular important to check for these things as the models' goodness of fit will depend on how well it predicts the dependent variable, the linearity of the model and the behaviour of the residuals. If the relationship between the independent variable(s) and the dependent variable seems to be curvilinear this study adds a square version of the variable, but if other characteristics than a linear or a U-shaped curve appears, the analysis of those variables will not continue any further. To examine the curvilinearity the command *scatter* is used to produce a scatterplot, which is a graphical representation of the correlation between the dependent variable and the independent variable(s).

An important assumption is that the variance in the residuals has to be homoskedastic[28] or constant (Stock and Watson, 2003, p. 126). To check for heteroskedacity in linear regression, it is needed to assess the residuals by fitted value plots. If a systematic change in the spread of the residuals over the range of measured values exist, it is most likely due to heteroskedacity in the dataset. If such pattern is visible, the model has a problem and the results may not be trustworthy.

To test for heteroskedacity in STATA, the *reg* command is used followed by the command, *rvfplot*.

[25] A curvilinear relationship is the same as a nonlinear relationship between two or more variables.
[26] "*Heteroskedacity is a statistics term that refers to the condition in which the variance of the error terms in a regression equation is not constant*" (Investiopedia, 2018)
[27] It is a phenomenon in which one independent variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy (Goldberger, 1991)
[28] Homoskedacity is a statistic term that refers to the event in which the variance of the errors in the dataset is similar (Investiopedia, 2018).

STATA will automatically draw a scatterplot between the residuals and the predicted values. If the residuals are not randomly distributed around the horizontal axis, it is an indication of heteroskedacity.

When running multiple linear regression analyses, it is important to check for multicollinearity. Voss (2004) differentiates between *full multicollinearity* and *partial multicollinearity*. Full multicollinearity is defined as: *When two or more explanatory variables overlap completely, with one a perfect linear function of the others, such that the method of analysis cannot distinguish them from each other"* (Voss, 2004, p. 1). Partial multicollinearity is defined as "*When two or more explanatory variables overlap, such that they are correlated with each other in a sample, but still contain independent variation. This condition limits the extent to which analysis can distinguish their causal importance, but does not violate any assumptions required for regression*" (Voss, 2004, p. 2).

To check for multicollinearity, this study uses the *vif* command in STATA right after running the regression. A *vif* value higher than 10 or a reciprocal *vif* value lower than 0.10 can indicate that the dataset is affected by multicollinearity. In such cases, the regression model has to be investigated further to conclude if the model is impacted by multicollinearity. By applying the Pearson Correlation Model, the *pwcorr* command, and the significance levels, command *sig*, in STATA, it is possible to evaluate the correlation between the independent variables and the significance levels for the correlations. Pearson correlation coefficients range from -1 to 1. The closer to ±1 the correlation is, the stronger the correlation and the probability for multicollinearity is high. However, if the significance coefficient is smaller than 0.95 (for confidence level = 95%), the probability for multicollinearity is limited.

To process to the second step of the analysis process, the P-value of the F-test, *F > prob* value (to be found in the *overall fit model*), has to be smaller than 0.05 (for the confidence level at 95%, which this study applies) as it means that the overall model may be significant. The third step in the analysis process is to examine the R^2 value and *adjusted* R^2 value[29]. If the R^2 value is higher or equal to 0.50 (with two significant digits) *and* the *adjusted* R^2 value is maximum 10% smaller than the R^2 value, the model will proceed to the fourth evaluation criterion. In the fourth step, the two tail p-value, *P>t*, is evaluated. The value is to be found in the *parameter estimate model*. If this value is lower than 0.05, there is a probability that the independent variable(s) is statistically significant in explaining the

[29] The adjusted R^2 is a transformed version of the R^2. The adjusted R^2 takes into account the number of predictors in the model (Miles, 2006). If the predictors in the model better the model, the adjusted R^2 increases.

dependent variable. One should be aware of the fact that when running multiple regression analyses not all independent variables are necessarily statistically significant in explaining the dependent variable. Only when all independent variables have significant impact on the dependent variable, the analysis will continue to the fifth step. In the last step of the analysis process, the model is tested for omitted variable biases. Missing variables in a model mean that the regression coefficients are inconsistent. The test for omitted variable bias is executed by using the Ramsey test, command *ovtest,* right after the regress command. If the model does not have omitted variable bias, the p-value is lower than the usual threshold of 0.05.

If and only if the linear regression model looks valid all the way through the five steps, an expert and the researcher evaluate the model. If the expert and the researcher evaluate the model as potentially having predictive power, the dependent variable $y = f(x_0)$ can be expressed as:

$$f(x_0) = b_0 x_0 + a$$

If and only if the multiple linear regression model looks valid all the way through the five steps, the model is a good predictor, and the dependent variable $y = f(x_n)$ can be expressed as:

$$f(x_{0\ldots}x_n) = b_0 x_0 + b_1 x_1 + b_2 x_2 \ldots b_n x_n + a$$

For both equations it applies that b_0, b_1, b_2 and b_n indicate the slope. The slope can be retrieved from the STATA *parameter estimate* output. Each slope value is stated in the *"coef"* column in the parameter estimate, and indicates the amount of change one could expect in dependent variable given a one-unit change in the value of that independent variable, given all other variables in the model are held constant. The sign of the regression coefficient indicates the relationship between the independent(s) and dependent variable ("-" indicates negative relationships and "+" indicates positive relationship), *a* is a constant and the indicates the *y*-intercept. The *y*-intercept can be retrieved from the STATA *parameter estimate* output as the "*_cons"*-value. It indicates the value of the dependent variable when the independent variable(s) equals zero.

Hereafter, the expert evaluates these models (see appendix 5). Lastly, due to the criticism of using social media for stock performance predictions emphasized by practitioners, this study has chosen to put a great effort on making sure that the probability of getting type I errors is as small as possible.

This is done by applying the most stringent multiple testing correction method, the Bonferroni correction method, and the least stringent correction method, the Benjamin-Hochberg False Discovery Rate to examine significant results.

5.2.7. *Trade ideas*

This study does not execute on the findings of the analysis, and it does not suggest investors whether or not they should trade findings of the study. Hence, the last step in J. P. Morgan's (2017) stepwise Big Data approach for investment professionals is irrelevant to answer the research questions, and is not further elaborated.

5.3. *Validity and Reliability*

Validity refers to the extent to which a study measures what it intends to measure and corresponds accurately to the real world (Kelly, 1927, p. 14). Validity is a requirement for all types of studies (Oliver, 2010). The methodology of the present study' should result in a high overall validity. High convergent validity is likely to be obtained if different combinations of methods lead to the same conclusion. Validity assessment of results and conclusions, is however, troubled, and the sampling methods are therefore relevant to focus on. The present study uses both qualitative and quantitative data collection methods to assess the investors' use of social media trading. In addition, the Bonferroni correction method and the Benjamini-Hockberg False Discovery Rate correction method are used in order to answer research question 3 and research question 4. High face validity can be obtained by subjectively reviewing how well the project is measuring what it intends to measure. To secure high face validity, this study has used pilot testing to adjust the interview protocol and the questionnaire to make sure that the design of the interviews and the questionnaire enable collection of qualified data to answer all research questions. Furthermore, the present study has used an expert to review the methods used and the results obtained in the secondary data analysis.

Reliability is the overall consistency of a measure (Trochim, 2006). It is a concern when the data source is *one* observer as it is hard to evaluate the observer's subjectivity (Babbie, 2010, p.158). Reliability issues are mainly linked to subjectivity (Wilson, 2010), but reliability is also very important to consider in quantitative research (Heale and Twycross, 2015). The high reliability of this study is achieved in several ways: first, telephone interviews are recorded and transcripted.

Second, the transcription is sent to each of the interviewees to make sure that the interview reflects their views correctly at the time of the interviews. Time-dependent changes in the outcome cannot be assessed by the methods used. A pilot test of one interview is performed to ensure ambiguity in the interview guidelines. To assure reliability of the questionnaire, three pilot tests are performed. Ambiguities, spelling mistakes and misunderstanding in the questions are sorted out and corrected. In addition, the study uses the same dependent and independent variables for the regression analyses performed per firm. By applying the same measurement tools, the reliability of the results increases (Trochim, 2006). To improve the reliability of this study, the secondary data analysis has to be repeated (see Recommendations for future research).

6. Data Analysis Process Diagram

The data analysis process diagram in Figure **11** pictures each step of the analysis methodology process for the secondary data in a sequential order. According to Tague (2004) "*the elements that may be included are: sequence of actions, materials or service entering or leaving the process (inputs and outputs), decisions that must be made, people who become involved and time involved at each step*".

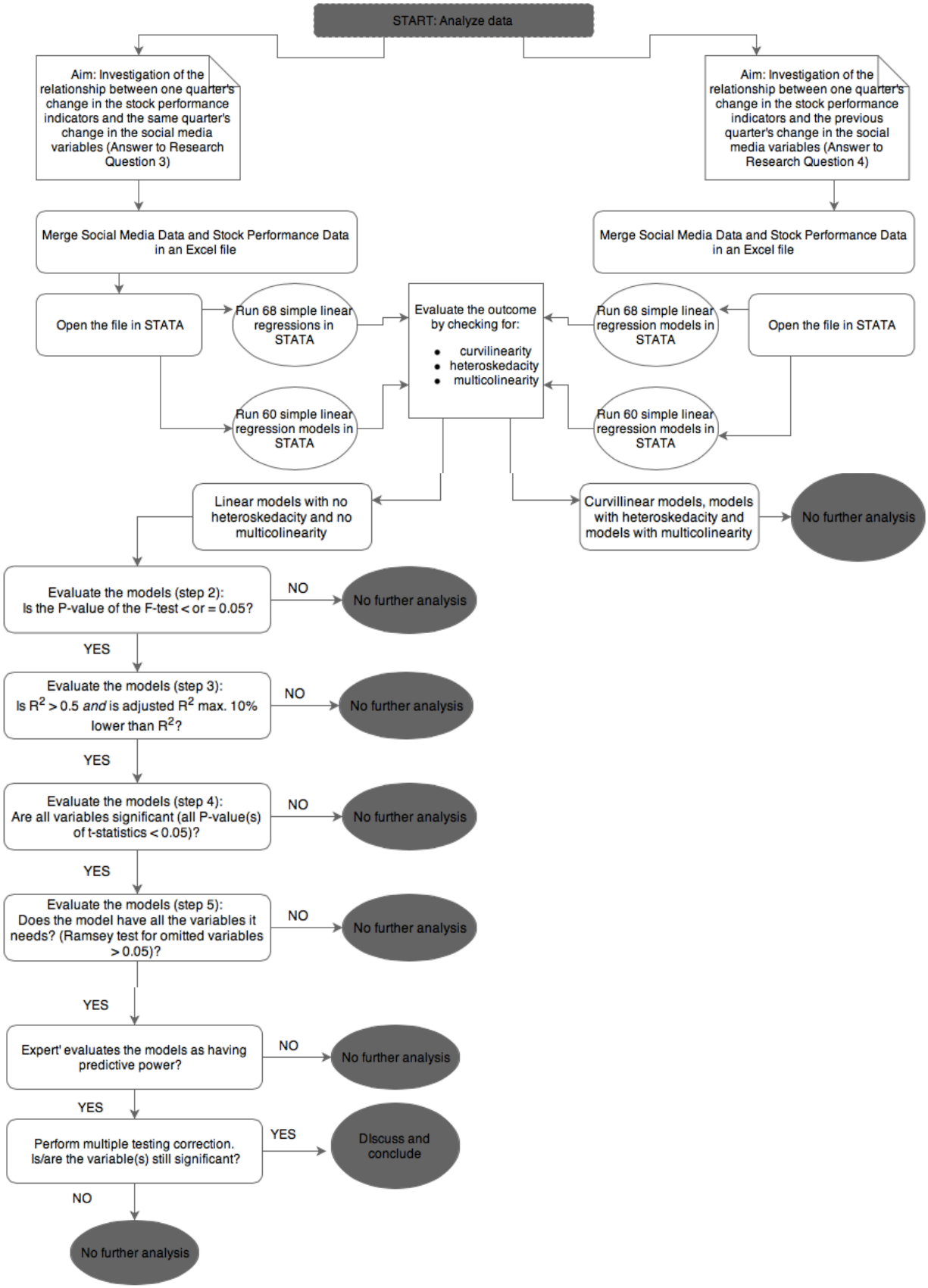

Figure 11: Process flow diagram: Analyze data

7. Summary of the collected data

In order for the reader to make sense of all the data that has been gathered in order to answer the research questions and how it relates to the AIDA model (awareness, interest, desire, action), it can be useful to refer to Figure **12**.

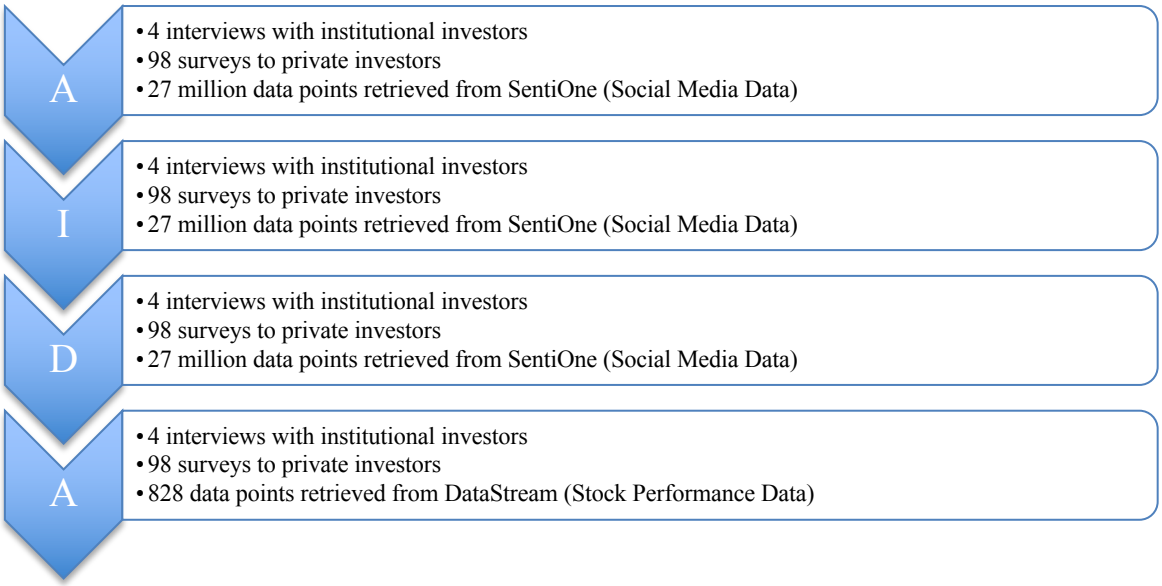

Figure 12: Summary of collected data

The figure helps the reader to understand the present study's data sources. In the next section it will be directly applied to the model (red. AIDA model) in order to answer the research questions.

8. Research Results

This section focuses on answering the research questions by applying the empirical data collected to each of the AIDA steps. The results section follows the structure of the AIDA model as it is divided into four parts: *awareness/attention, interest/knowledge, desire/preference* and *action/purchase*. The section is split into two overall parts for the matter of simplicity. The first section is the *primary data part*, where the interview results with professional investors and the questionnaire results with the private investors are presented. The second section is the *secondary data part*, which presents the research results from Social Media Data (covers A, I, D) and the Stock Performance Data (covers A). Moreover, the secondary data part presents results on the regression analysis between the Social Media Data and the Stock Performance Data, and it evaluates if the Social Media Data have any predictive power over the Stock Performance Data.

8.1. *Primary data*

To answer research question 1 and 2, it is determined to which degree the professional investors' and private investors' use social media trading as a trading strategy. As stated in the methodology section, the interviews and the surveys have an AIDA structure aiming to understand the investors' awareness about social media trading. In addition, the study aims to examine the degree of knowledge and interest within social media trading, the desire and preference towards social media trading over other trading strategies, and if the investors are currently using or are planning to use social media trading. The analyses of the primary data outline the research results for both professional investors and private investors.

8.1.1. *Awareness/attention*

The interviews conducted and surveys obtained are analyzed. It provides the possibility for this study to understand if the two kinds of investors have entered the first phase in the AIDA model based on their awareness and attention towards social media trading.

8.1.1.1. Interviews with professional investors

The degree of awareness of social media trading in a professional setting varied from one professional investor to another. Some interviewees are aware of the concept:

"I know that it exists, but I do not have any experiences with it myself (FV, 2018), and "*I know the phenomenon. I have also heard about some hedge funds that use it*" (PJ, 2018), whereas another only reported that he had heard about the concept: "*I have heard about it"* (AV, 2018)".

One professional investor is particularly aware of social media trading and mutual funds' usage of it:

"..Twitter is in particular experiencing a lot of interest by mutual funds based on artificial intelligence, which again is based on huge amounts of real time data from social media called 'big data'. In addition, social media listening can work as an "early warning" for some sectors where statistical material gets published with a major delay. This is pointed out by JP Morgan in 'Big Data and AI Strategies'"(FV, 2018).

Another investor is not aware of social media trading in a professional context at all:

"When it comes to professional investors..no, I have never heard about it..but I know that there are several Facebook groups, and probably other media as well, where stocks are debated..so there are surely some people that somehow attribute the social media content a value and use it in their strategy" (HD, 2018).

It is clear to the researcher that three out of four of the professional investors' awareness of social media trading in a professional manner is very limited. Even two of the professional investors initially are a bit confused about what social media trading in practice means, even though the definition of the concept of this study is included in the first part of the introduction e-mail sent by the researcher.

In summary, three (AV, PJ, FV) out of the four professional investors (AV, PJ, HD, FV) are aware of social media trading, but two have some difficulties in saying much more than that they have heard about it. In conclusion, 75% of the professional investors enter the first phase in the AIDA model.

8.1.1.2. Surveys to private investors

Out of the 98 surveyed, 56 respondents answer that they do not know about social media trading prior to their engagement in the survey. The remaining 42 respondents answer that they either pay great attention to the trading strategy or that they somehow are aware of it:

"*I was not aware of the name, but I am aware of the fact that one can partly or entirely base trading strategies on the basis of social media".*

Of the 42 respondents, three state that they became aware of the phenomenon through Aktieporteføljen for the first time between two and five years ago, and 11 respondents indicate that the first time they pay attention to social media trading is through Social Media like Facebook, Reddit or forums such as eToro, Euroinvestor and Proinvestor.

One investor state that, as a private investor, he/she is very aware of the information search and trying to grasp all information available on the Internet, including social media before investing.

In conclusion, most of the private investors are not aware of social media trading. However, the ones that are, have become aware of it through social media and online platforms. In summary, 43% out of the private investors (n=98) enter the first phase in the AIDA model.

8.1.2. *Interest/knowledge*

This section evaluates the professional and private investors' interest and knowledge about social media trading (second phase of the AIDA model).

8.1.2.1. Interviews with professional investors

Only two of the three professional investors that is aware of social media trading give indication of knowing about social media trading. One, in particular, shows higher degree of knowledge of the subject:

"*I do not know any institution that uses it, but everybody indeed wants to figure out to which extent social media can be applied to make more informed decision. Almost all professional investors use Bloomberg, and Bloomberg has also started to do social media listening. I know Bloomberg and Twitter have agreed to collaborate on the sentiments of tweets about specific stocks and how these are affected by the tweet sentiment. The second of April there was a fluctuation in Amazon's stock price due to a Trump post, and more information on Twitter since the inauguration of Trump has affected the stock prices. I do recommend you to read "Social Media Analytics: A survey of techniques, tools and platforms" by B. Batrinca, 2015"* (FV, 2018).

Another Chief Strategist also talks about his knowledge of social media trading:

"*I don't know if any firms use it here in Denmark, but maybe there are. However, I have heard about some algorithm funds that have been raised. I don't know the precise names of them, but I know that some of them collaborate with Google. Surely, this is to some extent related to social media trading, as they collect a lot information from online sources"* (PJ, 2018).

The remaining professional investor's knowledge and interest within social media trading is very limited:

"..it would be a remarkable strategy for professional investors to use social media for making trades..It is not something Danske Bank have looked further into" (AV, 2018).

In the interview with Danske Bank, the researcher tells the interviewee that some American hedge funds use social media trading. At first, the interviewee does not sound very excited in response to that information, but after considerations, he says:

"*Well yes, but there are almost hedge funds for everything today… Danske Bank offers a broad selection of stocks and funds, including hedge funds, so I cannot say that we do not use social media trading indirectly"* (AV, 2018).

The interesting possibilities social media trading could offer a company are outlined by some of the interviewees:

"*When you do not look on a single website, blog or Facebook site, but it is the sum of all social media, it is interesting which insights you can get out of it, I think..it is an interesting topic, and the more normal it becomes for the financial industry to listen to what is said on social media, the more money are invested into it"* (PJ, 2018).

"*It is super interesting and it is great that we get some knowledge about it, because it urges us to look more into it and consider if it is something we should use* (AV, 2018).

PJ (2018) further elaborates on his previous comment:

"*..there is a huge difference of knowing that a stock is traded and that you read about a stock can be traded..social media trading is a bit like Tripadvisor: you don't know if the raters are making fake suggestions..in addition, there is a risk that robots just create and publish a lot of posts that in consequence can affect your investment decision if you exclusively look at the content that is posted on social media"*

One professional investor seems to have more knowledge of the phenomenon than the others and argues:

"I am confident that there is no documentation on the ability to create abnormal returns by using social media trading. A study from 2011 stated that with the accuracy of 88% one could predict the coming two days flucations of the Dow Jones…the data was used by the hedge fund Derwent Capital Markets which within a year got filed for bankruptcy" (FV, 2018).

In conclusion, two (PJ, FV) out of the three professional investors (AV, PJ, FV) that are aware of social media trading also show interest in it. AV does not indicate much knowledge about the phenomenon, but during the interview, his interest in social media trading and its application increases a bit. In summary, 50% of the professional investors (n=4) enter the second phase in the AIDA model.

8.1.2.2. Surveys to private investors

The respondents are asked if they have searched or looked for information on social media trading before. Out of the 42 private investors that are aware of social media trading, 9 investors indicate that the concept is of such high interest to them that they have looked for more information about the trading method. 10 private investors state that they cannot remember if they have searched for more information about the topic, and the remaining 23 private investors answer 'no' to that question.

The respondents are also asked if they have heard about anyone who uses social media trading as a trading strategy. 33 respondents have not heard about anyone who uses social media trading. The same nine persons as the ones who have looked for more information about social media trading answer 'yes' to this question. Two of the nine private investors list a few names on private investors who use social media trading: "*Adel Hussain"*, "*I know Carl Icahn, whom I follow on Facebook, is using it"*. Two of the nine private investors state that they know "many people" who use it. One respondent answers that he knows of several American funds using this trading strategy. One respondent knows "a few people" who uses social media trading as a secondary trading strategy, and that those people put a lot of effort into it due to high risks. One respondent uses it to get a feeling of the market psychology of the stocks of his/her interest.

In conclusion, out of the 42 private investors who are aware of social media trading, nine of them are also interested in the phenomenon and show knowledge of social media trading. In summary, 9.2% of the private investors (n=98) qualify to enter the second phase in the model.

8.1.3. *Desire / preference*

The investors' desire and preferences towards social media are evaluated to investigate if they qualify to enter the third phase in the AIDA model.

8.1.3.1. Interviews with professional investors

The interviewees who are both aware *and* interested in social media trading indicated that they had limited preferences towards it:

"No, we do not prefer this strategy over fundamentals. It is too risky. In addition to what I said before about the study of 2011, it is not an appropriate strategy to manage a global portfolio of stocks. The amount of information is too big and therefore, you will need robots to filter and register the information..if we should have a preference towards social media trading, some requirements need to be in place. For instance, you need to be able to separate fake news from 'real news' and also a special knowledge about every single source behind the content posted on social media" (FV, 2018).

"*..if you decide on your next trading based on what is written on social media, you do not know if that trade fits your risk profile"* (PJ, 2018).

FV also states that to get social media trading to work for investors, one needs the information before other investors. If everybody uses this strategy, no benefit of 'knowing' first would exist, and no abnormal returns could be made.

The investors agree on the fact that social media trading does not indicate fundamentals of a stock, which is crucial to know about when investing and therefore doing such analysis is preferred over social media trading:

"*..we use fundamentals in our analysis where we look at i.e. cash flow..and other factors that have to be in place such as the global BNP..if social media trading should be used it should be in combination with some of those other strategies and fundamental analysis"* (PJ, 2018)".

"Social media trading is missing the fundamentals. It is typically all the research you get on a business, an economy or political research that you have to pay for. It is usual that a subscription to get that news can cost hundreds of thousands of dollars and it is often investors who buy that information. Of course they keep it for themselves. The information that is posted on social media will only be tiny parts of the entire in-depth analysis. You don't get the full picture through social media" (FV, 2018).

However, one of the two investors does see potential in using some kind of social media signal:

"*..if you can develop an algorithm that ensures the spreading of risk and you, as a professional investor, know what you do, and if only the minority of the firms on social media uses robots..you can maybe get return based on that strategy"* (PJ, 2018).

The same professional investor highlighted even other applications of social media listening:

"*..as a risk management tool, it could be used to track negative mentions about companies on social media and on that basis give each stock a risk profile. I really see the potential of this, because that works as a quantitative warning signal"* (PJ, 2018)

In contrast to PJ, FV stated that social media trading would contradict the vision and mission of the firm he worked at due to Pension Denmark's principles of investment horizons:

"*I want to highlight the fact that pensions funds hold a very long time investment horizon, which will not be possible with social media trading"* (FV, 2018).

Even though one (PJ) out of the two professional investors who are aware of and have interest in social media trading sees potential in the usage of social media as a warning monitor, this study consideres both professional investors' desire to include it as a trading strategy as not existing. Their preferences are to use fundamentals, micro indicators and macro indicators to making trade decisions.

In conclusion, none of the professional investors, 0%, have a desire to use social media trading.

8.1.3.2. Surveys to private investors

The respondents are asked about their preference towards social media trading. Out of the nine private investors who enter both the first *and* second phase of the AIDA model, six state that they do prefer other trading strategies to social media trading as they do not expect any economic benefit with this type or because their principles of thorough stock analysis conflict with the principles used in social media trading analyses. This is due to a few reasons:

"*I don't trust people on social media"*, *"there are really stupid people on social media, I will not gamble with my money on what they say"*, and *"I always prefer fundamentals"*.

The remaining three respondents state that they have a preference towards social media trading because they know the economic benefit associated with this type of trading strategy, and they think *'it is a great tool to determine the optimism or pessimism in the market'*. However, none of these three private investors would use social media trading as the only trading strategy. They would rather combine social media trading with fundamentals analysis.

In conclusion, out of the nine private investors who enter the previous phases, 3 private investors have a desire/preference for using social media trading as a part of their investment strategy. In summary, three private investors (3%) enter the third phase in the AIDA model.

8.1.4. *Action/purchase*

The investors' actions on and purchases of social media trading are evaluated to investigate if they qualified to enter the fourth phase of the AIDA model.

8.1.4.1. Interviews with professional investors

The professional investors are asked if they use social media trading or if their organization have any plans to use it in the future.

According to the planned questions the researcher can now have asked whether the professional investors use social media trading or if their organization plans using it. However, as none of the professional investors have a desire/preference to use social media trading in their organization, the professional investors, could, according to theory, not enter the fourth phase in the AIDA model.

One professional investor makes an interesting comment when asked if the firm he is employed with has any plans on the adaptation of social media trading:

"*..the banking sector is rather conservative, so I will, however, assume that it should come from the outside if the banking sector should be reformed in this field*" (PJ, 2018).

His statement implies that innovation of the banking sector would probably not arise from internal sources. In relation to that, it could be relevant to look at fin-tech start-ups and ask them the same questions.

8.1.4.2. Surveys to private investors

Out of the three private investors who enter the third phase, all state that they already use social media trading, and they will continue to use the trading strategy:

"*Yes, it is an active part of my strategy..but in combination with other strategies*" . Another private investor answered that: *"I use it because of the opportunity for an abnormal return"*, and the last private investor stated: *"I have used it for index trading based on a technical analysis made on social media content".*

In conclusion, the three private investors that enter the previous phase, all use social media trading. Hence, these 3 respondents show action on social media trading. In summary, 3% of the private investors (n=98) also enter the fourth phase in the AIDA model.

8.2. *Sub-conclusion on research question 1*

Based on the interviews conducted with the professional investors, it is concluded that one out of four is not aware of social media trading, and thus, does not enter the first phase (awareness of/attention to social media trading) in the AIDA model. Out of the three professional investors who enter the first phase, two also qualify for the second phase in the AIDA model (interest in/knowledge about social media trading). However, neither of those two professional investors have a desire to use social media trading nor preferences to using it with other trading strategies. They both prefer fundamentals and global micro analyses and global macro analyses. Hence, none of the investors enter the third phase (desire for/preference towards social media trading) of the AIDA model, and do therefore not qualify for the fourth phase (action on/purchase social media trading) either. Figure **13** sums up the findings.

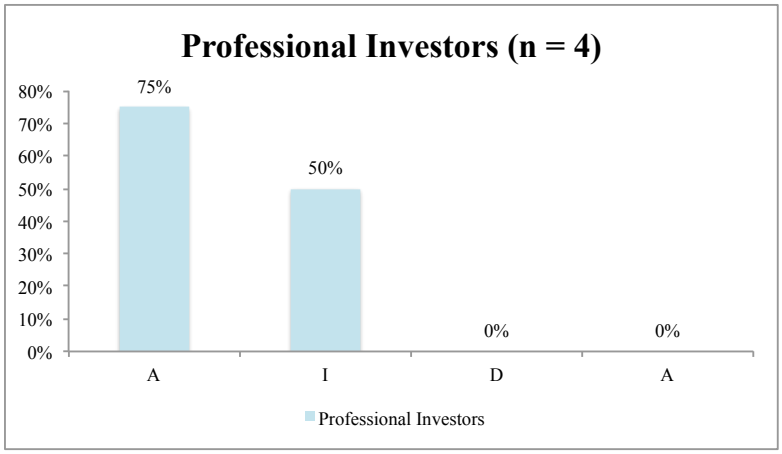

Figure 13: Professional investors that have entered the specific phases of the AIDA model (%)

8.3. *Sub-conclusion on research question 2*

When it comes to the private investors, the responses are more varied and dispersed than the professional investors. Most of the respondents are not aware of the social media trading strategy. Out of the private investors who are aware of social media trading (42 out of the 98 responses), only nine of them have an interest in it to a degree where they gain knowledge about the trading strategy. Out of those nine private investors who enter the second phase in the AIDA model, six do not have a desire/preference towards social media trading, and thus, they do not enter the third phase in the AIDA model. However, the remaining three both have a desire/preference towards social media trading and the same three private investors already use it. Hence, these 3 private

investors have both entered phase 3 and phase 4 in the AIDA model. Figure **14** sums up the findings.

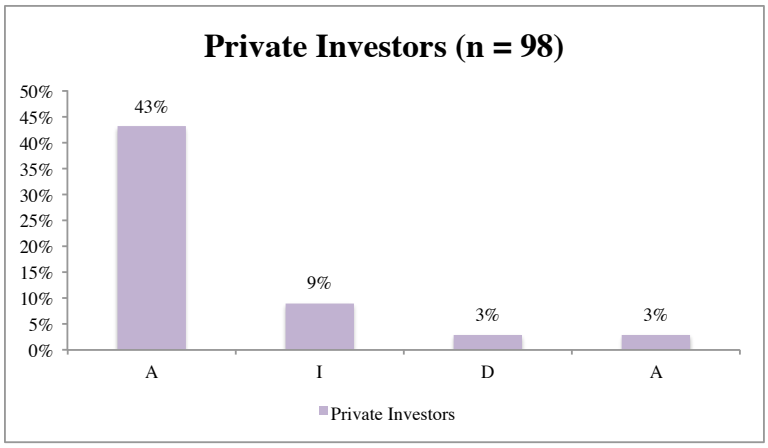

Figure 14: Private investors that have entered the specific phases of the AIDA model (%)

8.4. *Secondary data results*

The findings of the secondary data research are presented in this section. These findings will help to answer research question 3 and research question 4. It should be clear to the reader that the social media research results measure the *awareness,* the *interest* and the *desire* for the products/services/stock of the five companies (examined in section *Social Media Data research results*) and the Stock Performance Data measure the degree of stock purchase *action* (examined in section *Stock Performance Data research results*) To answer research question 3 and 4, the relationship between the '*AID'* and the '*A'* needed to be investigated through regression analysis. It is done in the section *Social Media Data and Stock Performance Data regression results.* This part of the secondary data results is the most comprehensive of all the three parts due to its importance to the conclusion.

8.4.1. *Social Media Data research results: awareness/attention, interest/knowledge and desire/preferences*

The level of awareness of, interest in and desire for Bavarian Nordic's, Coloplast's, Danske Bank's, Genmab's and Jyske Bank's products/services/stocks on social media are determined by the identification, acquisition, storing, structuring and pre-processing of social media mentions as described in the methodology section.

Activities on social media about all five firms are observed in all quarters from second quarter in 2015 (Q215) to the fourth quarter in 2017 (Q417). From the observations, it is concluded that users on social media are *aware*, *interested* and have a *desire* towards all the five firms' stock or/and its products or/and services.

With 52,099 mentions on social media, Jyske Bank is the firm that gets the most attention from social media users. Danske Bank is the company with second highest mentions on social media. In the concerned period, Danske Bank is mentioned 31,123. Genmab receives 27,187 mentions. Coloplast is mentioned 27,054 times and Bavarian Nordic is mentioned 26,760 times.

Most articles arewritten about Genmab and Bavarian Nordic, and most posts are created about Jyske Bank. Jyske Bank is also the firm with the highest number of positive sentiments. Danske Bank ranks highest on the number of negative sentiments, and 99% of these negative sentiments are created on Facebook. Facebook is, without comparison, the domain where most mentions about the five companies are posted. It is followed by mentions on portals and mentions on Twitter. Jyske Bank is both the one with the largest amount of content created on Facebook and the firm with the

most positive sentiments in the Facebook posts. Most of the photos and videos created are about Coloplast, whereas most blogs and tweets concern Bavarian Nordic.

The period with highest social media activity for the five firms is in Q216 and Q117. In total, the firms are mentioned 29,745 times in Q216, and 27,984 times in Q117 (it is 37% of all mentions from Q215-Q417).

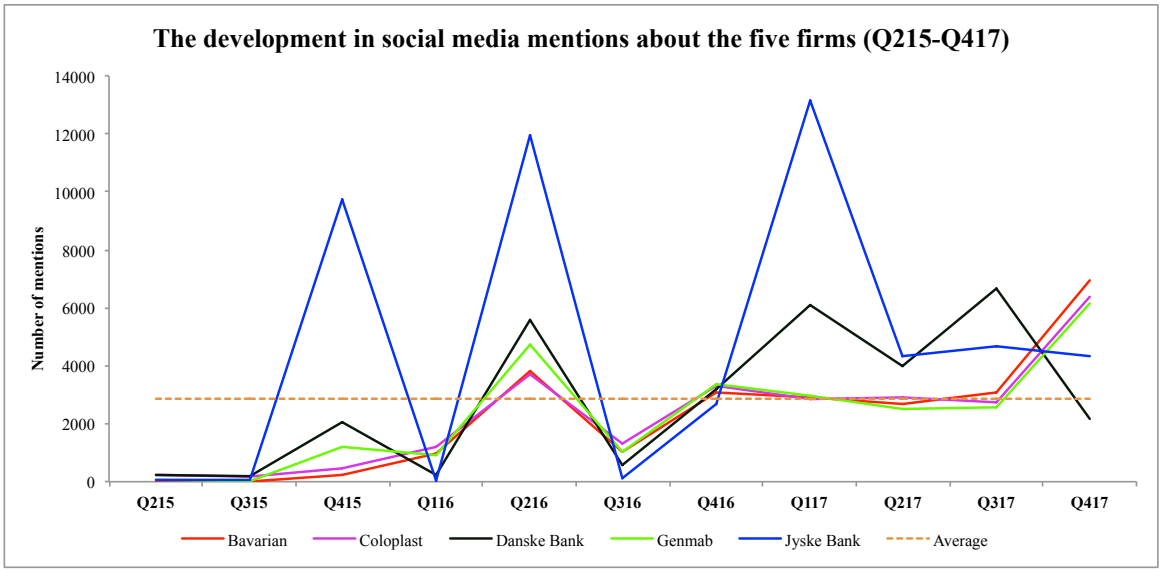

Figure 15: The development in social media mentions about the five firms (Q215-Q417)

From Figure **15** it is seen that there is an increasing number of mentions about the five firms over the quarters. The reason for this could be due to the spread of social media in Denmark, but also the spread of corporate usage of social media as a communication tool. When the companies create accounts on social media it incites social media users to contact and comment more on those companies (Hoffman and Fodor, 2010).

The following three figures (Figure **16**, Figure **17** and Figure **18**) depict the change of the five firms in three of the social media variables out of 19 social media variables that are used for the regression analysis.

Figure **16** pictures the quarterly change in the number of mentions about each of the five firms on Facebook. The *y*-axis explains the percentage change. Figure **16** shows that the number of quarterly mentions about each of the five firms on Facebook increased from Q116.

Danske Bank and Jyske Bank experience the highest quarterly changes in the amount of mentions on Facebook. The changes of these two firms in Facebook activity follow the same pattern from Q315 to Q117. Apart from the two Danish banks, Danske Bank and Jyske Bank, the mentions on

Facebook about the three other firms look rather stable with increases and decreases between 10% and -0.5%.

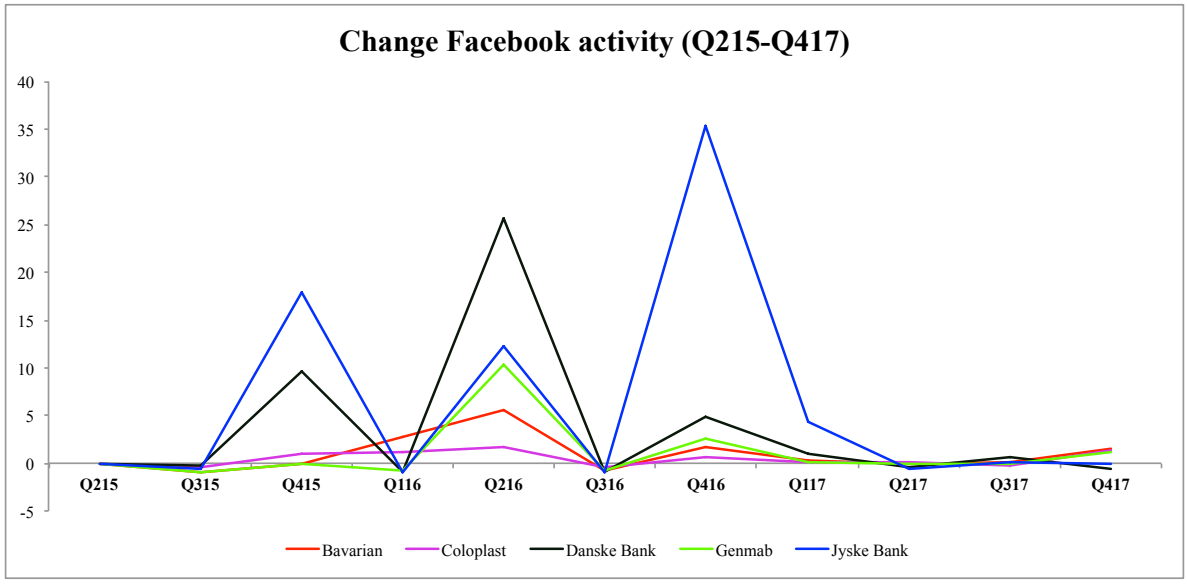

Figure 16: Change in mentions on Facebook about the five firms (Q215-Q417)

The second most used social media type to write about the five firms is the *portals*. Figure **17** depicts the quarterly percentages change in the mentions about the five firms on portals in the period of Q215 to end of Q417.

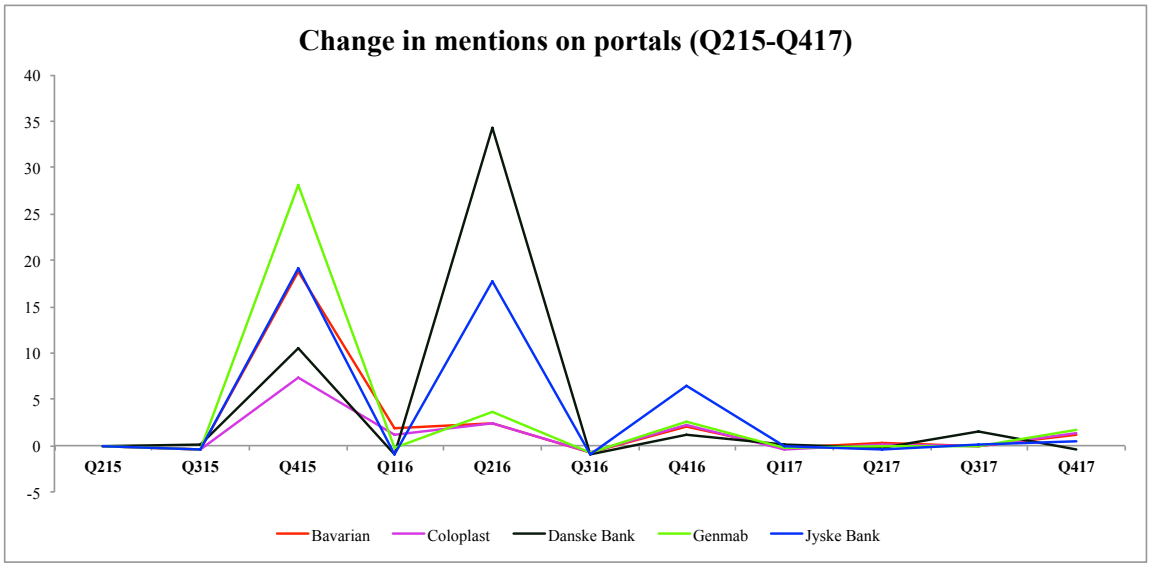

Figure 17: Change in mentions on portals about the five firms (Q215-Q417)

It is observed that the change in the mentions on the portals about the firms follows the same pattern from Q315 to Q217. The highest percentage increase from one quarter to another is from Q116 to Q216 where mentions about Danske Bank on portals increase with about 35%. The percentage changes in the mentions on portals reach a stable level after Q117.

The third most used social media type is Twitter. The graphs below picture the quarterly percentages change in the mentions on Twitter in the period end of Q215 to Q417. Genmab is the firm that receives most attention on Twitter and the firm with the highest percentage increase (almost 90% increase) from Q315 to Q415. From Genmab's annual report (2015-2016), it is stated that the firm generates 60% higher revenue in 2015 compared to the year before. This positive financial development could possibly give rise to more debate about Genmab on media, including Twitter and help explain the spike in Twitter activity in Q4-15 (cf. Figure **18**).

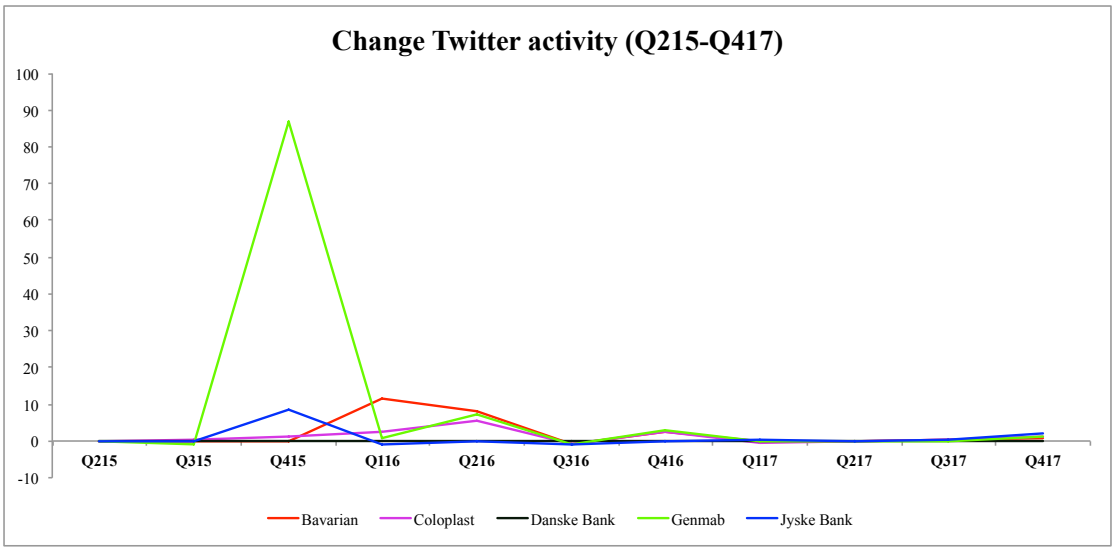

Figure 18: Change in mentions on Twitter about the five firms (Q215-Q417)

A slight pattern is seen in Figure **18**. It indicates that more mentions were published about the five firms in the period from Q415 to Q216. From Q216, the percentage change in the mentions about the five firms on Twitter remains almost unchanged.

8.4.2. *Stock Performance research results: Action/purchase*

For each company, 828 Stock Performance Data points are collected through DataStream in an Excel file. The data points include the weekly stocks' opening price, the volume traded, the P/E value, and the P/B value from the second quarter of 2015 (Q215) to the last quarter of 2017

(Q417). Based on that data, the change in opening price, volume traded, P/E value, and P/B value are calculated on a weekly and a quarterly basis. Figure **19** envisions the quarterly change in the opening price for each of the five firms.

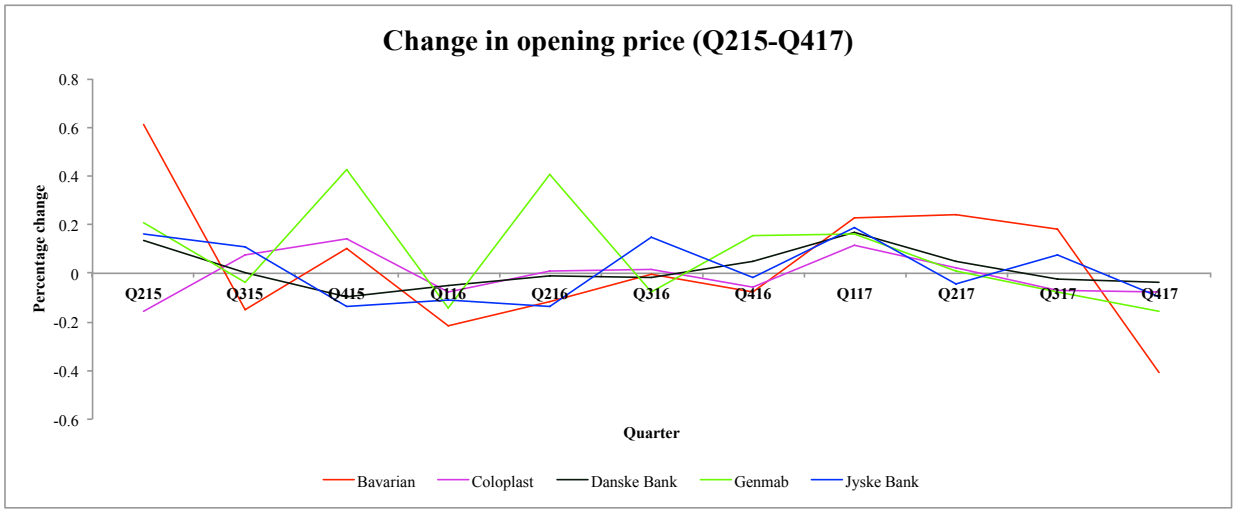

Figure 19: Quarterly increase/decrease (%) in the opening prices (Q215-Q417)

It is seen on Figure **19** that all the five firms' opening prices fluctuate. There does not seem to be any correlation between the five firm's opening price percentage increase/decrease. The percentage changes range from 0.6 to -0.4. Bavarian Nordic and Genmab are the two firms experiencing the highest percentage increases and decreases in their opening price. The stocks' fluctuations of these firms can be caused by the firms' relatively narrow product portfolios and uncertainty of future products. The biotechnology industry is more risky (Nisen, 2017). It is therefore not a surprise that the two biotechnology stocks experience higher fluctuations.

When looking for some trends in the volume traded of the five stocks (see Figure **20**), it seems like there is a clear pattern in the last couple of quarters (a substantial increase followed by a substantial decrease for all stocks). Also, in most of the other quarters, the increases and decreases in the volume traded are parallel. The Bavarian Nordic stock is the stock with the largest percentage increase (+1.2%). The largest percentage decrease happens in Danske Bank in Q315, where the volume of Danske Bank stocks decreases with about 0.4% compared to the previous quarter.

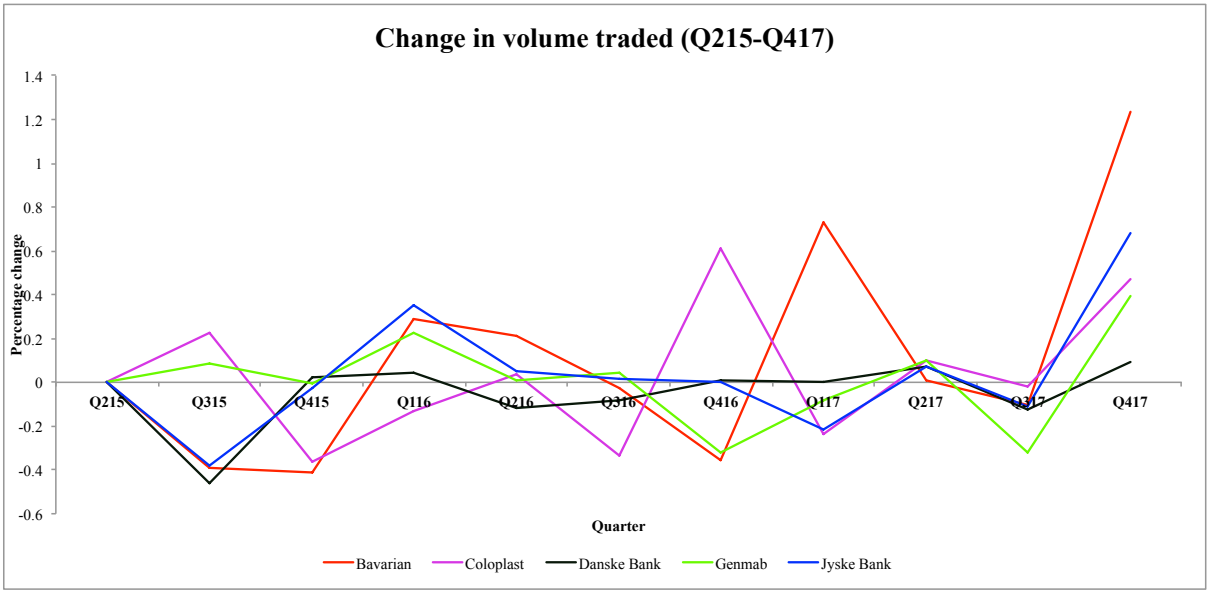

Figure 20: Quarterly increase/decrease (%) in the volume of stocks traded (Q215-Q417)

When it comes to the percentage changes in the firms' P/E ratio, it becomes clear that no general trends for the five firms existed. This can be due to the very firm-specific earnings and unparalleled changes in the firms' stock prices. On Figure **21** it is visible that the percentage change in the P/E ratio isonly ranging from 3% and -0.5%, so the P/E ratio is rather stable for all the firms. Coloplast is the firm with the highest percentage change in its' P/E ratio (Q415). All stocks' P/E ratio developments are rather stable, but the two biotechnology companies, Genmab and Bavarian Nordic are experiencing the highest variance in the percentage change looking at the full period, whereas the percentage change in the P/E ratio of the mature firms, Jyske Bank and Danske Bank, are more stable. The explanation to that could be that the product portfolios of these firms are more diversified, and thus the future state of these two firms does not depend on how well a single or a small number of products perform, which means that their stock price is less likely to fluctuate as much as the price of Bavarian Nordic and Genmab.

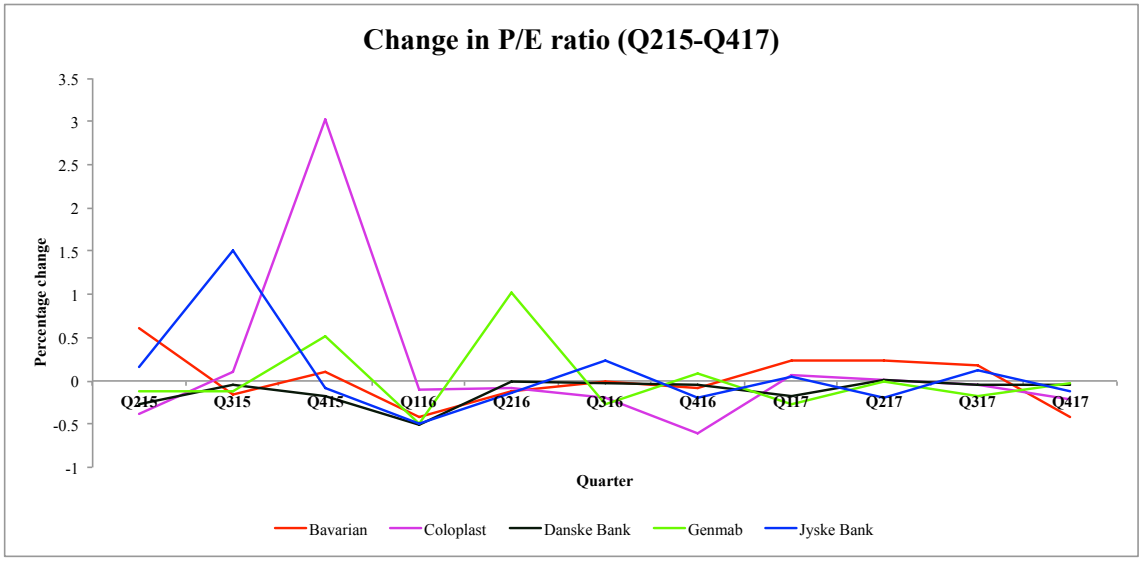

Figure 21: Quarterly increase/decrease (%) in P/E ratios (Q215-Q417)

The last stock performance indicator evaluated is the percentage change in the P/B ratio. As for the stock performance indicator evaluated just before, the percentage increase/decrease is very small for all the firms (between +0.4 and -0.6).

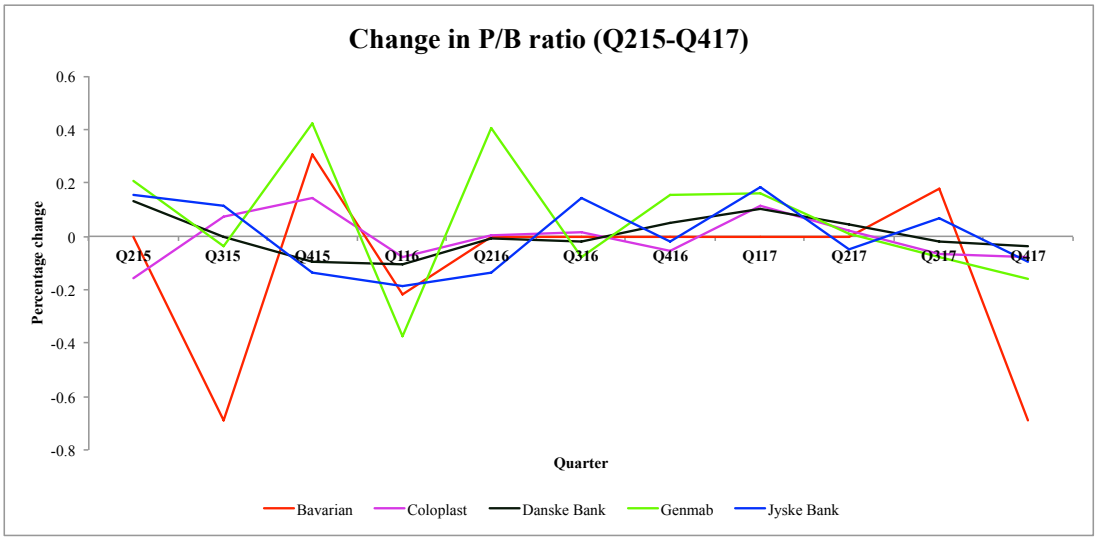

Figure 22: Quarterly increase/decrease (%) in P/B ratios (Q215-Q417)

The characteristics of the graphs pictured in Figure **22** indicate that the volatility of the biotechnology firms' (Bavarian Nordic and Genmab) P/B ratios is much higher than the three other firms' P/B ratio volatility, which are more constant. This might be due to the biotechnology firms' product portfolio being smaller and more dependent on few products. It creates uncertainty in the

expectations towards the firms' performances, thus the higher stock price variation volatility in the P/B ratio.

8.4.3. *Social Media Data and Stock Performance Data regression results*

This section presents the outcome of the regression analysis. In total 272 (76*2+ 60*2) analyses are performed per firm, and eight different dependent variables are used. The table below sums up the outcome of the analyses by specifying which dependent variables that potentially could (+) and could not (-) be significantly correlated with one or more independent variables.

Dependent variable	At least one correlation (+/-)	Related to Research Question (3 or 4)
$\Delta OpeningPrice_{t+1}$	+	3
$\Delta P/E_{t+1}$	+	3
$\Delta P/B_{t+1}$	-	3
$\Delta \sum Volume\ traded_{t+1}$	-	3
$\Delta OpeningPrice_{t+2}$	+	4
$\Delta P/E_{t+2}$	+	4
$\Delta P/B_{t+2}$	-	4
$\Delta \sum Volume\ traded_{t+2}$	-	4

Figure 23: Summary of dependent variable characteristics from the regression models

Figure **23** shows that none of the changes in the social media variables are correlated with the change in the P/B ratio in the same or following quarter. Further to that, none of the changes in social media variables are correlated with the change in the volume traded in the same quarter or the next quarter. Figure **24** summarizes the 19 different independent variables used for the simple and multiple regression analyses. Each independent variable is applied in 76 simple linear regression models. The column *'#regression analyses performed with x'* specifies the total number of regression models performed per firm in which the independent variable is used as a predictor. In consequence, if the independent variable is only used in simple linear regression models, the number in the column will be 76. Oppositely, if the "*#regression analyses performed with x"* indicates that i.e. 88 regression analyses have been performed, the total number of multiple linear regression analyses performed is 12 (88-76 = 12), and the number of simple linear regression analyses

performed is 76. The combinations of the independent variables used for the multiple linear regression analyses are visible in appendix 11.

Independent variable(x)	#regression analyses performed with x	#potentially statistically significant models x is a part of
$\Delta\sum articles_t$	88	4
$\Delta\sum posts_t$	88	5
$\Delta\sum mentions\ on\ portals_t$	84	4
$\Delta\sum mentions\ on\ Facebook_t$	88	0
$\Delta\sum mentions\ on\ Twitter_t$	88	1
$\Delta\sum mentions\ on\ Blogs_t$	84	0
$\Delta\sum mentions\ on\ Forums_t$	84	0
$\Delta\sum$mentions on Reviews	80	0
$\Delta\sum mentions\ to\ Photo\&Video_t$	84	0
$\Delta\sum positive\ sentiments_t$	100	5
$\Delta\sum negative\ sentiments_t$	100	6
$\Delta\sum positive\ mentions\ on\ Facebook_t$	76	0
$\Delta\sum negative\ mentions\ on\ Facebook_t$	76	0
$\Delta\sum positive\ mentions\ on\ Twitter_t$	76	0
$\Delta\sum negative\ mentions\ on\ Twitter_t$	76	1
$\Delta\sum positive\ mentions\ on\ portals_t$	76	1
$\Delta\sum negative\ mentions\ on\ portals_t$	76	0

Figure 24: Summary of independent variable characteristics from the regression models

As is shown in Figure **24**, eight out of 19 independent variables used for this study are correlated with at least one dependent variable. The number of regressions performed is at a level making it

irrelevant to consider a type II error. In conclusion, mentions about the five C20 firms published on Facebook, blogs, forums, reviews, Instagram or Youtube do not impact the change in any stock performance variables. In addition, positive or negative mentions on Facebook do not impact the stock performance indicators, nor do the positive mentions on Twitter or the negative mentions on portals of the five firms.

Figure **25** presents the simple and multiple regression models that have the possibility of being statistically significant.

Model	Firm	Dependent variable (f(x))	Independent variable (x_0)	Independent variable (x_1)	R^2 value	P-value(s) F-statistics	Formula ($f(x)$)
A	Coloplast	$\Delta P/E_{t+1}$	$\Delta\sum articles_{t+1}$	$\Delta\sum positive\ sentiments_{t+1}$	0.7635	0.0121	$f(x_0,x_1)=0.481x_0-0.574x_1-0.071$
B	Genmab	$\Delta OpeningPrice_{t+1}$	$\Delta\sum articles_{t+1}$		0.5810	0.0002	$f(x_0)=0.064x_0-0.042$
C	Genmab	$\Delta OpeningPrice_{t+1}$	$\Delta\sum posts_{t+1}$		0.5523	0.014	$f(x_0)=0.091x_0-0.030$
D	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum mentions\ published\ on\ portals_{t+1}$		0.7037	0.002	$f(x_0)=0.223x_0-0.162$
E	Genmab	$\Delta OpeningPrice_{t+1}$	$\Delta\sum posts_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.7635	0.0031	$f(x_0,x_1)=0.004x_0+0.058x_1-0.039$
F	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum posts_{t+1}$	$\Delta\sum positive\ sentiments_{t+1}$	0.8117	0.0013	$f(x_0,x_1)=0.006x_0+0.143x_1-0.152$
G	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum articles_{t+1}$	$\Delta\sum positive\ sentiments_{t+1}$	0.8150	0.0012	$f(x_0,x_1)=0.021x_0+0.129x_1-0.156$
H	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum tweets_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.8441	0.0006	$f(x_0,x_1)=0.008x_0+0.145x_1-0.166$
I	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum articles_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.8485	0.0005	$f(x_0,x_1)=0.021x_0+0.136x_1-0.169$
J	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum mentions\ published\ on\ portals_{t+1}$	$\Delta\sum positive\ sentiments_{t+1}$	0.8102	0.0013	$f(x_0,x_1)=0.041x_0+0.132x_1-0.154$
K	Genmab	$\Delta P/E_{t+1}$	$\Delta\sum mentions\ published\ on\ portals_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.8441	0.0006	$f(x_0,x_1)=0.024x_0+0.139x_1-0.167$
L	Bavarian	$\Delta P/E_{t+2}$	$\Delta\sum negative\ sentiments\ published\ on\ Twitter_{t+1}$		0.7373	0.001	$f(x_0)=0.041x_0+0.030$
M	Danske Bank	$\Delta P/E_{t+2}$	$\Delta\sum positive\ sentiments\ published\ on\ portals_{t+1}$		0.7747	0.001	$f(x_0)=-0.039x_0+0.063$
N	Danske Bank	$\Delta OpeningPrice_{t+2}$	$\Delta\sum mentions\ published\ in\ portals_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.7626	0.0065	$f(x_0,x_1)=-0.100x_0+0.046x_1-0.009$
O	Jyske Bank	$\Delta OpeningPrice_{t+2}$	$\Delta\sum posts_{t+1}$	$\Delta\sum positive\ sentiments_{t+1}$	0.6529	0.0246	$f(x_0,x_1)=0.197x_0-0.009x_1-0.052$
P	Jyske Bank	$\Delta P/E_{t+2}$	$\Delta\sum posts_{t+1}$	$\Delta\sum negative\ sentiments_{t+1}$	0.7425	0.0087	$f(x_0,x_1)=0.001X_0-0.002X_1-0.069$

Figure 25: Summary of the models (A-P) that potentially could be statistically significant

All five firms studied are represented in Figure **25**. Hence, at least one of each of the five firms' stock performance indicators are potentially correlated with the social media variables. Out of the 272 regression models that are performed per firm, (1360 regressions in total for all five firms), 16 regression models could potentially be statistically significant. Out of the 16 regression models, 11 are significant when correcting for multiple comparisons (see below).

8.4.3.1. Multiple testing corrections

There are different kinds of multiple testing corrections. This study uses the most conservative, *the Bonferroni correction method*, and a less stringent correction, *the Benjamini-Hochberg False Discovery Rate Method*.

- *The Bonferroni Correction Method*

The Bonferroni correction of all p-values is made. The critical value *a* is set at a 0.05 level. As eight different dependent variables are used, and the number of tests performed on each of the eight dependent variables varies, the corrected critical value varies. For each of the dependent variables, 15 multiple regression analyses are performed (n_m = 15). Model *A, E, F, G, H, I, J, K, N, O, P* areall multiple regression functions, and the corrected critical value a_c is therefore calculated to be:

$$\frac{\alpha}{n_m} = \alpha_c \rightarrow \frac{0.05}{15} = 0.0033$$

For the model to be significant, the corrected p-value, p_i, has to be less than 0.05 to be significant at the 5% level when 15 analyses are made. However, there is still a 5% risk that the study does not reject an insignificant model leading to a type I error.

For each of the dependent variables, 19 simple regression analyses are performed (n_s = 19). Model *B, C, D, L, M* are all simple regression functions, and the corrected critical value α_c is therefore calculated to be:

$$\frac{\alpha}{n_s} = \alpha_c \rightarrow \frac{0.05}{19} = 0.0026$$

For the model to be significant, the corrected p-value, p_i, has to be less than 0.05 to be significant at the 5% level when 19 analyses are made. However, there is still a 5% risk that the study does not reject an insignificant model, and thus, making a type I error.

The Bonferroni corrected P-values, p_i, are calculated with the following formula

$$p_i = p * n_s \; OR \;\; p_i = p * n_m$$

If $p_i < \alpha$, the model is significant after the Bonferroni correction has been made. Each model's significance after the Bonferroni correction is made, is stated in the column '*Significant?*' in Figure **26**.

Model ref	**P-value (*p*)**	**Corrected critical value** α_c	**Bonferroni corrected P-value (p_i)**	**Significant?**
A	0.0121	0.0033	0.1815	-
B	0.0002	0.0026	0.0038	+
C	0.014	0.0026	0.266	-
D	0.002	0.0026	0.038	+
E	0.0031	0.0033	0.0465	+
F	0.0013	0.0033	0.0195	+
G	0.0012	0.0033	0.018	+
H	0.0006	0.0033	0.009	+
I	0.0005	0.0033	0.0075	+
J	0.0013	0.0033	0.0195	+
K	0.0006	0.0033	0.009	+
L	0.001	0.0026	0.019	+
M	0.001	0.0026	0.019	+
N	0.0065	0.0033	0.0975	-
O	0.0246	0.0033	0.369	-
P	0.0087	0.0033	0.1305	-

Figure 26: Bonferroni Correction Results

From Figure **26**, it is clear that five models (*A, C, N, O, P*) are not significant after the Bonferroni correction. Due to the conservatism of the Bonferroni correction method, the Benjamini-Hochberg False Discovery method is also performed.

- *The Benjamini-Hochberg False Discovery Rate Method*

The Benjamini-Hochberg correction of the model's p-values, p, are applied. The false discovery rate (FDR) a is set at a 0.05 level. The models' (*A-P*) p-values are ranked in order, from the smallest (0.0002) to the largest (0.0246). The largest p-value is multiplied with the number of models that are significant (n = 16). The second largest p-value (0.014) is thereafter multiplied with the number of models in the test divided by its rank (r =1) to find the Corrected Benjamini-Hockberg P-value p_j:

$$p_j = p * \frac{n}{n - r}$$

The third largest p-value has rank r = 2, and so forth. If the Corrected Benjamini-Hockberg P-value p_j is smaller than a, the model is significant according to the Benjamini-Hockberg theory.

Figure **27** presents the results of the Benjamini-Hockberg corrections. Importantly, both the FDR_α = 0.05 and FDR_α = 0.01 and FDR_α = 0.005 are applied. By using the FDR_α = 0.05, the probability to make a type I error is less than 5%. By applying the FDR_α = 0.010, the probability to make a type I error is less than 1%. By applying the FDR_α = 0.005, the probability to make a type I error is less than 0.5%. Most studies for larger datasets use FDR_α = 0.05, whereas studies of smaller datasets justify the application of FDR_α = 0.01. The FDR_α = 0.005 is rather conservative.

Model	P-value	Benjamini-Hockberg	Significant at	Significant at	Significant at

ref		Corrected P-value (p_j)	FDR_α =0.05?	FDR_α =0.01?	FDR_α =0.005?
A	0.0121	0.0138	+	-	-
B	0.0002	0.0023	+	+	+
C	0.014	0.0149	+	-	-
D	0.002	0.0032	+	+	+
E	0.0031	0.0045	+	+	+
F	0.0013	0.0023	+	+	+
G	0.0012	0.0023	+	+	+
H	0.0006	0.0023	+	+	+
I	0.0005	0.0023	+	+	+
J	0.0013	0.0023	+	+	+
K	0.0006	0.0023	+	+	+
L	0.001	0.0023	+	+	+
M	0.001	0.0023	+	+	+
N	0.0065	0.0086	+	+	-
O	0.0246	0.0246	+	-	-
P	0.0087	0.0107	+	-	-

Figure 27: Benjamini-Hockberg Correction results

The results of the Benjamini-Hockberg corrections presented in Figure **27** above yield very different results depending on the FDR_α used. At $FDR_\alpha = 0.05$, all models are significant. At $FDR_\alpha = 0.01$, four models are not significant. These are: *A, C, O, P*. At $FDR_\alpha = 0.005$ the models *A, C, N, O, P* are not significant.

- *Comparison of Bonferroni Correction results and Benjamini-Hochberg Correction results*

By comparing the two correction methods, it isevident from the analysis, that at the same critical value $\alpha = 0.05$, eleven models (*B, D, E, F, G, H, I, J, K, L, M*) are, according to the Bonferroni Correction, still significant and five models (*A, C, N, O, P*) are not. In contrast, the Benjamini-Hockberg Correction method does reject that all models are not significant at that α-value. However, by limiting the probability of making type I errors from 5% to 1% (which is still a high

probability compared to the ones used for the Bonferroni correction), the Benjamini-Hochberg Correction method then rejects four models (*A, C, O, P*) to be significant using the $FDR_{\alpha} = 0.05$, the probability to make a type I error is less than 5%. By applying the $FDR_{\alpha} = 0.010$, the probability to make a type I error is less than 1%. At the $FDR_{\alpha} = 0.005$, the stringency of the Benjamini-Hockberg correction method is almost at the same level as for the Bonferroni correction method. At this level, the results of the two analyses yield the same results: Therefore, it is concluded that the following models are statistically significant: *B, D, E, F, G, H, I, J, K, L, M* even when corrected for multiple comparisons.

The following section will comment on the firm specific results obtained from the analysis conducted.

8.4.3.2. Firm specific patterns in the results

The three firms' (Genmab, Danske Bank and Bavarian Nordic) statistically significant models *(B, D, E, F, G, H, I, J, K, L, M*) are discussed in the three following subsections.

- *Genmab*

In total, nine significant models describing the change in Genmab's stock performance indicators and the change in social media mentions are found. Two out of the nine models with predictive power are associated with the change in the opening price, and seven models see a correlation between the change in the P/E ratio and changes in social media indicators. Notably, most of the independent variables are recurrent in more regressions. In fact, the only independent variable that only occurs once is *the change in the sum of tweets.* The second independent variable x_1 is in all models either the change in the sum of negative sentiments or the change in the sum of positive sentiments. Hereby, it can be concluded that the type of sentiment is not relevant to the change in Genmab's stock performance indicator.

Moreover, in four out of the ten statistically significant models, the *change in the sum of mentions about Genmab posted on portals* is an independent variable; in three out of the ten statistically significant models' the *change in the sum of articles about Genmab* isan independent variable; in three out of the ten statistically significant modelsthe *change in the sum of posts about Genmab* is an independent variable. It is therefore concluded that Genmab's changes in its stock performance indicators are not driven by sentiments in social media, (apart from what is posted on the portals) but rather by the total sum of all the mentions of the firm on social media (cf. the *Conceptual Framework* all mentions about the firms

are either grouped into 'posts' or articles). Therefore, it is not very important on *which* social media Genmab is mentioned; the important thing is that there *are* mentions. One more interesting discovery based on the table above is that all constants (a) are negative and that all the slopes (b_0, b_1) are positive (see Figure **25**). Hereby, it is concluded that the higher the change in the social media independent variables the higher is the increase in Genmab's stock performance indicators. The following figures below present each of the observed (actual) values and the predicted values of the nine Genmab models. It is notable that the value of the slope in model D is much higher compared to the other significant models. This suggests that mentions published about Genmab on portals play a more significant role as compared to the other independent variables.

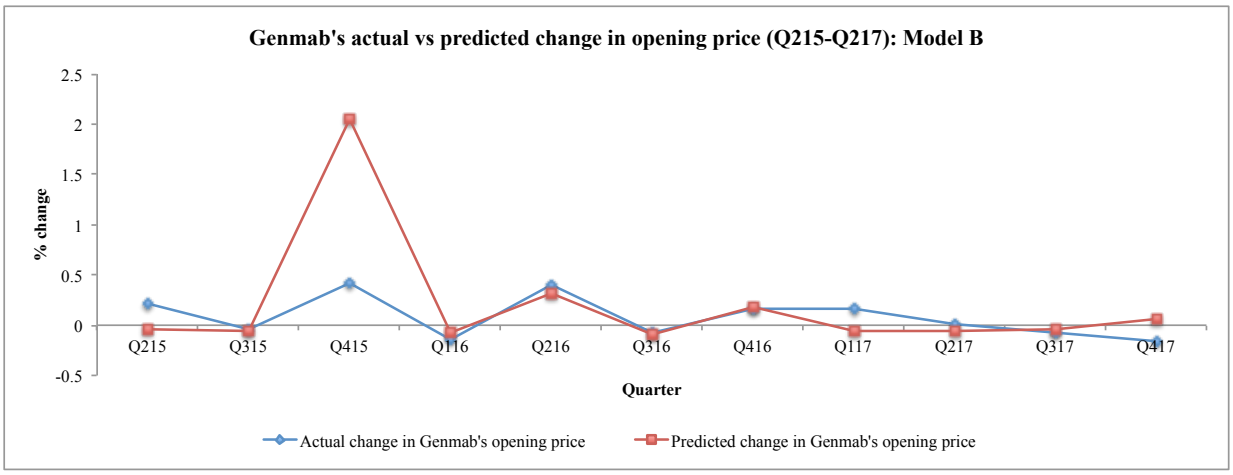

Figure 28: Model B: Genmab's actual vs. predicted change in opening price (Q215-Q417).

The data shown in Figure **28** illustrates the curves for actual changes in Genmab's opening price (blue line) and values for the predicted changes (red line) in Genmab's opening price ratio for each quarter from the second quarter in 2015 to the last quarter in 2017. It is evident from Figure **28** that the actual changes and the predicted changes follow each other nicely with only one exception in Q415.

In this case the dependent variable is the change in opening price (ΔOpening price$_{t+1}$) and the independent variable is the change in the number of articles posted about Genmab ($\Delta\sum$articles$_{+1}$) The predictive model is written as (see Figure **25**):

(B) $$f(x_0) = 0.064x_0 - 0.042$$

The formula should be interpreted as a one-unit increase in $\Delta\sum$arrticles$_{+1}$ lead to a 6.4% increase in Genmab's ΔOpening price$_{+1}$. In consequence, 6.4% of Genmab's change in its opening price could be explained by the change in the number of articles published about Genmab. However, the

imprecise prediction in Q415 could imply that the predictive model is lacking accuracy, as the change in articles about Genmab from one quarter to another is higher than 1.5%.

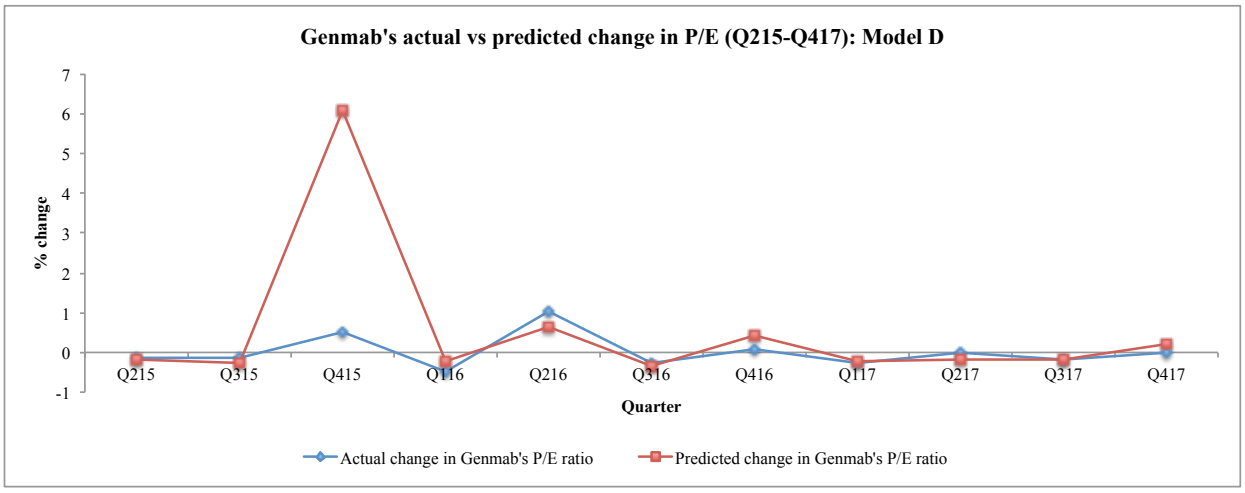

Figure 29: Model D: Genmab's actual vs. predicted change P/E ratio (Q215-Q417).

The data shown in Figure **29** pictures the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E ratio for each quarter from the second quarter in 2015 to the last quarter in 2017. It is obvious from Figure **29** that the actual changes and the predicted changes follow each other nicely with only one exception in Q415, as seen in Model B.

In this case the dependent variable is the change in P/E ratio ($\Delta P/E_{t+1}$) and the independent variable is the change in mentions about Genmab posted on portals ($\Delta\sum$mentions published on portals$_{t+1}$). The predictive model is:

(D) $$f(x_0) = 0.223x_0 - 0.162$$

The formula should be interpreted as a one-unit increase in $\Delta\sum$mentions published on portals$_{t+1}$ leads to a 22.3% increase in Genmab's $\Delta P/E_{t+1}$. It means that 22.3% of Genmab's change in its P/E ratio is explained by the change in the number of mentions it receives on portals. However, from Figure **29**, Q415, it could be interpreted that for high values of x_0, the predictive model could have some limitations.

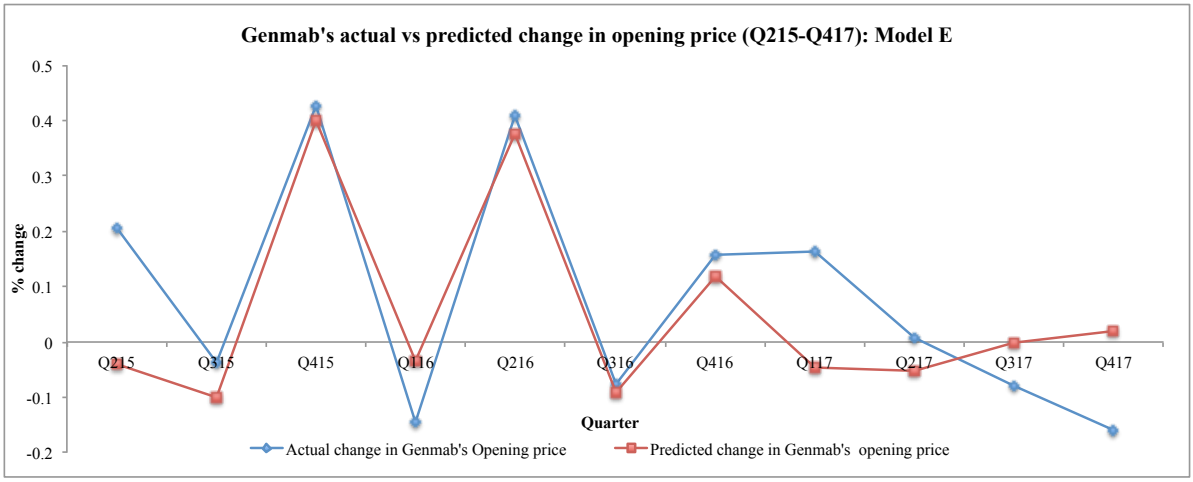

Figure 30: Model E: Genmab's actual vs. predicted change in opening price (Q215-Q417).

The data shown in Figure **30** showed the curves for actual changes in Genmab's opening price (blue line) and values for the predicted changes (red line) in Genmab's opening price for each quarter from the second quarter in 2015 to the last quarter in 2017. It is obvious from the data that the actual changes and the predicted changes follow each other quite nicely apart from Q215, Q117 and Q417 even though a lot of fluctuations in the firm's change in its opening price from quarter to quarter take place. In this case the dependent variable is $\Delta\text{Openingprice}_{t+1}$ and the two independent variables are the change in posts ($\Delta\sum\text{posts}_{t+1}$) and the change in negative sentiments in the posts ($\Delta\sum\text{negative sentiments}_{t+1}$). The predictive model is:

(E) $$f(x_0, x_1) = 0.004x_0 + 0.058x_1 - 0.039$$

The formula should be interpreted as a one-unit increase in $\Delta\sum\text{posts}_{t+1}$ leads to a 4% increase in $\Delta\text{Openingprice}_{t+1}$. Likewise, a one-unit increase in $\Delta\sum\text{negative sentiments}_{t+1}$ leads to a 5.8% increase in $\Delta\text{Openingprice}_{t+1}$.

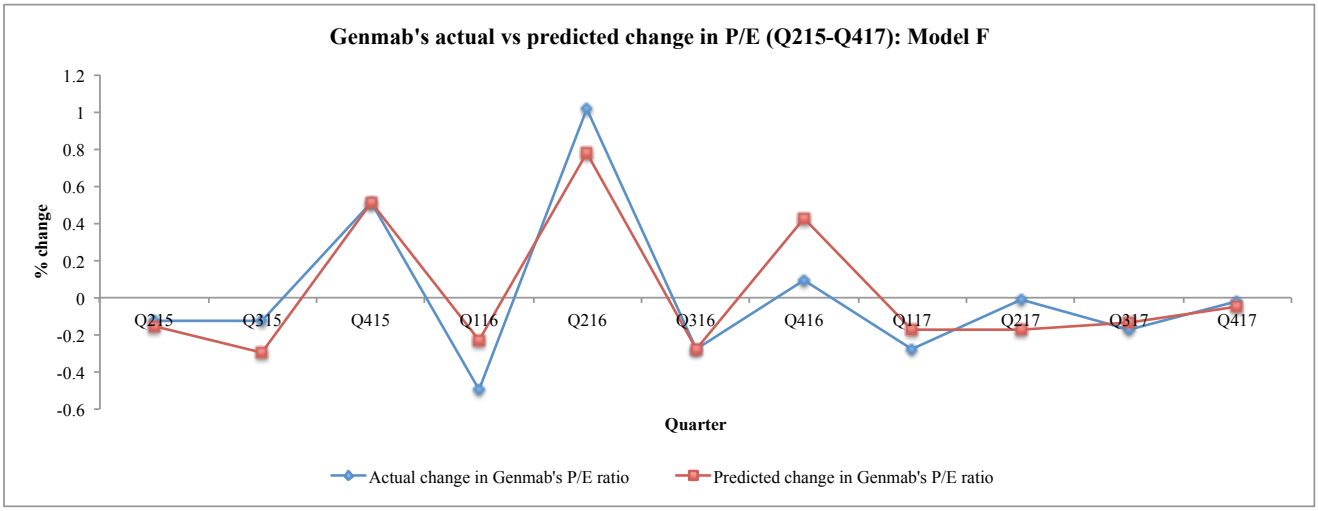

Figure 31: Model F: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

The data shown in Figure **31** show the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E ratio for each quarter from the second quarter in 2015 to the last quarter in 2017. In this case the dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in posts ($\Delta\sum posts_{t+1}$) and the change in positive sentiments in the posts ($\Delta\sum positive\ sentiments_{t+1}$). The model is:

(F) $$f(x_0, x_1) = 0.006x_0 + 0.143x_1 - 0.152$$

It is obvious from the data that the actual changes and the predicted changes followed each other to a large extent even though some fluctuations from quarter to quarter take place.

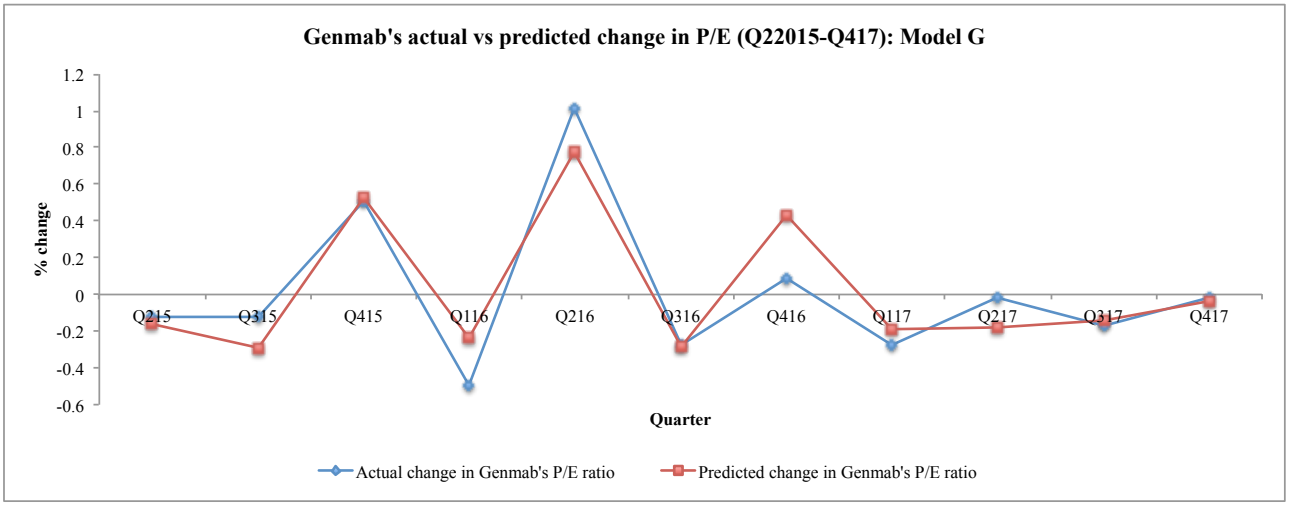

Figure 32: Model G: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

The data presented in Figure **32** shows the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E for each quarter from Q215 to Q417. Again, the actual changes and the predicted changes followed each other well in that period. In this case the dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in articles ($\Delta\sum articles_{t+1}$) and the change in positive sentiments in the posts ($\Delta\sum positive\ sentiments_{t+1}$). The model is:

(G) $$f(x_0, x_1) = 0.021x_0 + 0.129x_1 - 0.156$$

According to the G model, 2.1% of a $\Delta P/E_{t+1}$ increase can be explained by a one-unit increase in $\Delta\sum articles_{t+1}$. Likewise, a one-unit increase in $\Delta\sum positive\ sentiments_{+1}$ leads to a 12.9% increase in $\Delta P/E_{t+1}$.

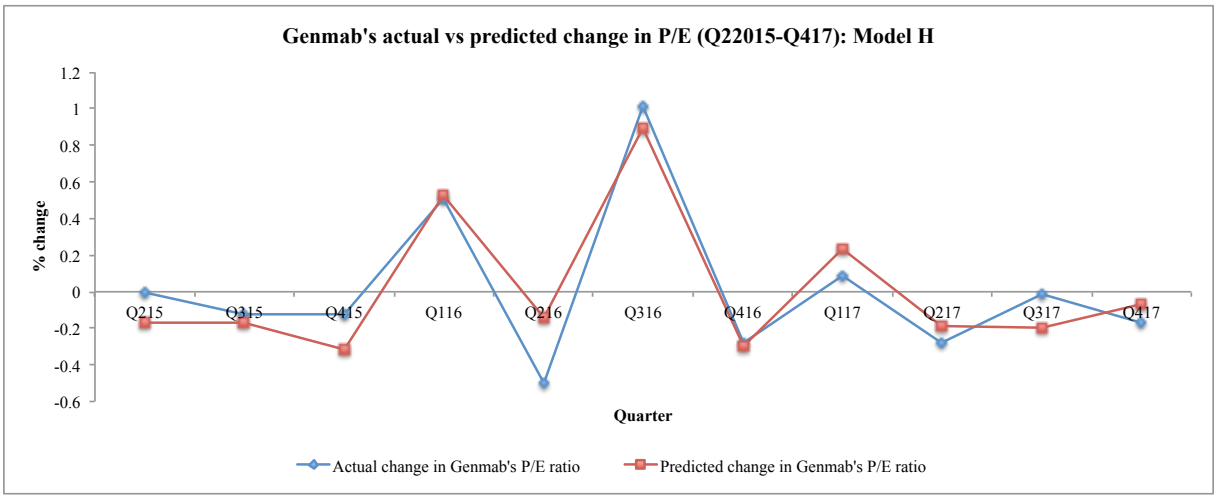

Figure 33: Model H: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

The data shown in Figure **33** portray the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E for each quarter from Q215 to Q417. As other Genmab figures, the two curves followed each other nicely. In this case the dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in tweets written about Genmab ($\Delta\sum tweets_{t+1}$) and the change in negative sentiments in the posts ($\Delta\sum negative\ sentiments_{t+1}$). The predictive model is:

(H) $$f(x_0, x_1) = 0.008x_0 + 0.145x_1 - 0.166$$

The formula should be interpreted as a one-unit increase in $\Delta\sum tweets_{t+1}$ explains less than 1% of increase $\Delta P/E_{t+1}$. Even though this independent variable is significant, it is arguable that it is not important to consider when determining factors that impact fluctuations in Genmab's P/E ratio. In contrast, a one unit increase in $\Delta\sum negative\ sentiments_{+1}$ lead to a 14.5% increase in $\Delta P/E_{t+1}$. This effect is indeed much larger and is therefore important.

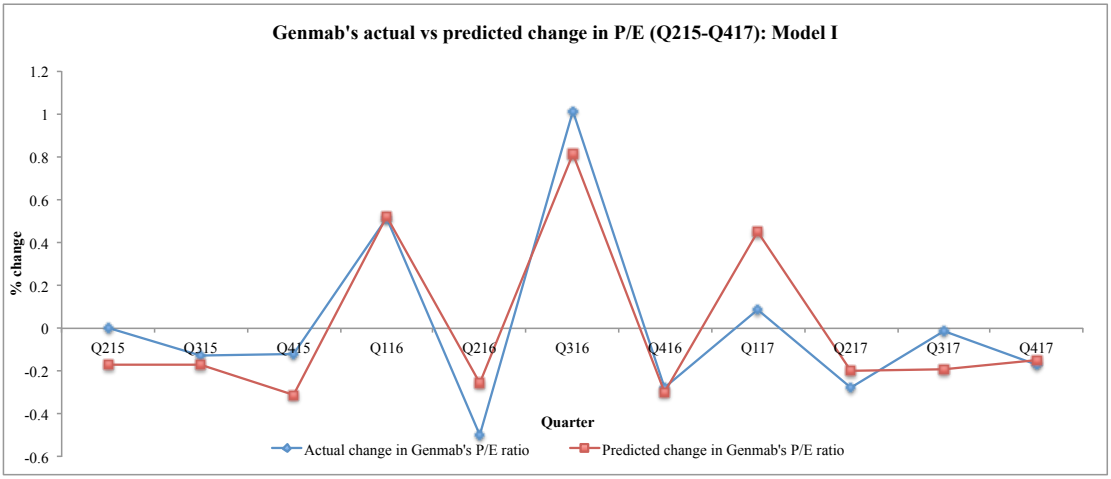

Figure 34: Model I: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

Figure **34** shows the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E ratio for each quarter from Q215 to Q417. The dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in articles ($\Delta\sum articles_{t+1}$) and the change in negative sentiments in the posts ($\Delta\sum negative\ sentiments_{t+1}$). Figure **34** shows that the actual changes and the predicted changes follow each other quite nicely. The predictive model can be written as:

(I) $$f(x_0, x_1) = 0.021x_0 + 0.136x_1 - 0.169$$

The regression equation implies that the change in P/E_{t+1} would increase by 2.1% if the change in articles increases with one-unit. Likewise, if the change in negative sentiment increases with one unit, the change in Genmab's P/E would increase with 13.6%. As in the model H case, it is relevant to discuss if an impact with 2.1% is important to consider.

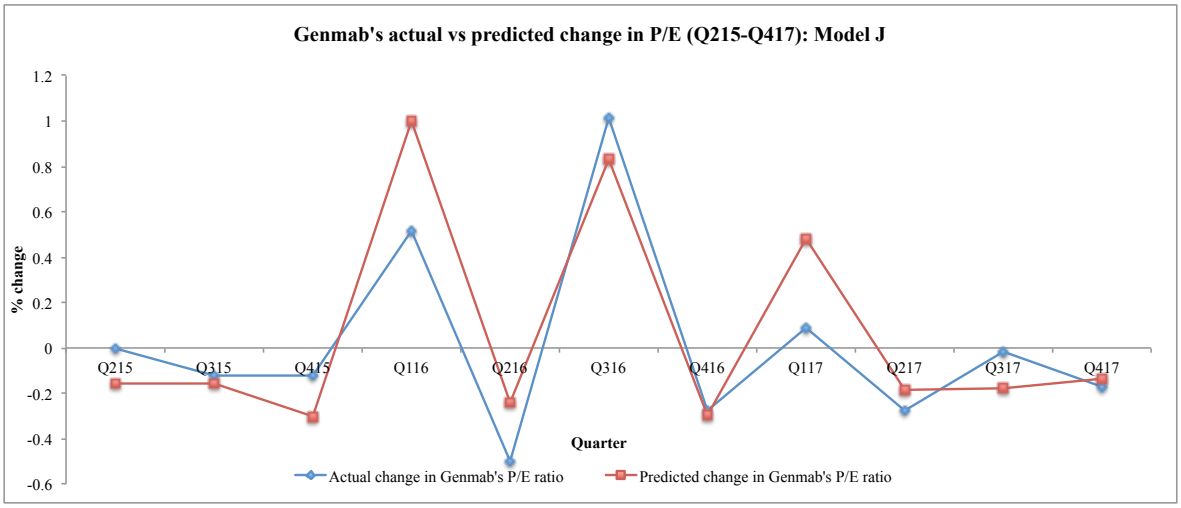

Figure 35: Model J: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

Figure **35** also shows the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E ratio for each quarter from Q215 to Q417. The dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in mentions posted on portals ($\Delta\sum$mentions on portals$_{t+1}$) and the change in positive sentiments in the posts ($\Delta\sum$positive sentiments$_{t+1}$). Compared to the previous models, the actual changes and the predicted changes do not follow each other as nicely as in other cases (i.e. *F, H, I*). The predictive model J is written as:

(J) $$f(x_0, x_1) = 0.041x_0 + 0.132x_1 - 0.15$$

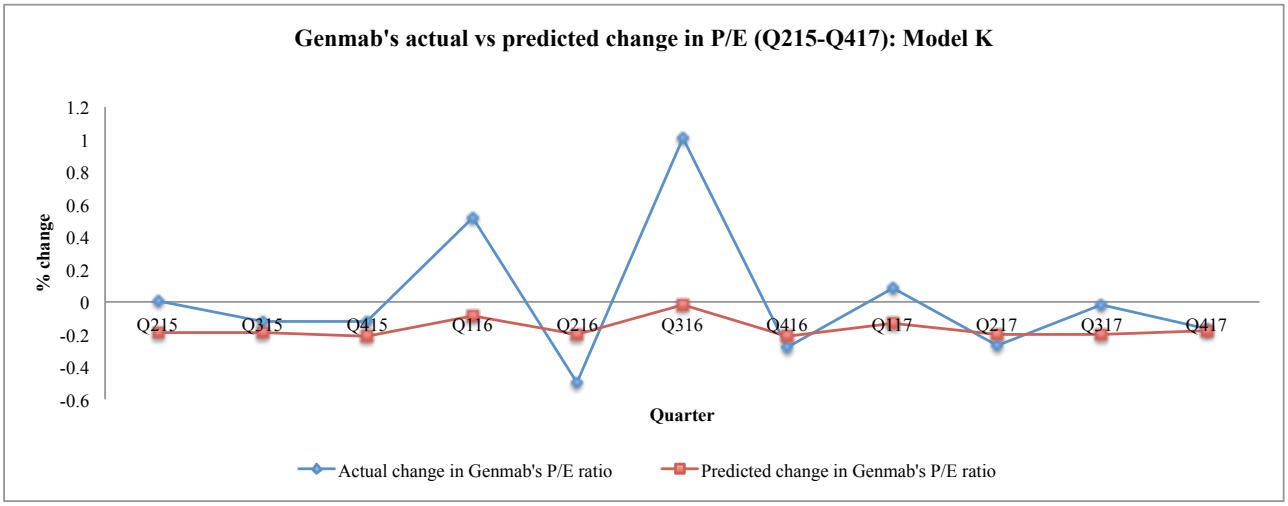

Figure 36: Model K: Genmab's actual vs. predicted change in P/E ratio (Q215-Q417).

Finally, Figure **36**, shows the curves for actual changes in Genmab's P/E ratio (blue line) and values for the predicted changes (red line) in Genmab's P/E ratio for each quarter from Q215 to Q417. The dependent variable is $\Delta P/E_{t+1}$ and the two independent variables are the change in mentions posted on portals ($\Delta\sum$mentions on portals$_{t+1}$) and the change in negative sentiments in the posts ($\Delta\sum$negative sentiments$_{t+1}$). This model resembles the previous model (model J). In fact, both the dependent variable and one of the two independent variables are similar. What distinguishes the two models (J and K) from each other is the last independent variable, x_1. However, by comparing the two figures (Figure **35** and Figure **36**), the actual change and the predicted change in Figure **35** follow each other to a larger extent. The predictive model K is:

(K) $$f(x_0, x_1) = 0.024x_0 + 0.119x_1 - 0.169$$

- *Danske Bank*

One statistically significant model has predictive power. The model is a simple linear regression model with one independent variable (ref. *M*) that can predict the change in Danske Bank's P/E ratio in one quarter based on the change in the total sum of positive sentiments published on portals. Derived from Figure **25** the mathematical expression is:

(M) $$f(x_0) = -0.039x_0 + 0.063$$

The slope of the regression line is negative. This suggests that positive sentiments published on portals are correlated to a reduction in Danske Bank's P/E ratio. If this reflects a true cause, and given that a reduction in the P/E reflects a fall in stock price, it could suggest that positive sentiments on portals induce stockowners to sell their stocks. Although this could seem counterintuitive, the same inverse relationship is observed in the Bavarian Nordic stock analysis (see Figure **38**). Figure **37** presents the observed (actual) values of Danske Bank's change in the P/E ratio the quarter after the change in positive sentiments about Danske Bank is published on the different types of portals. For the individual quarters, the model and the actual changes do not follow each other very well. The model predicts a rather constant value, and the actual development in P/E ratio varies considerably.

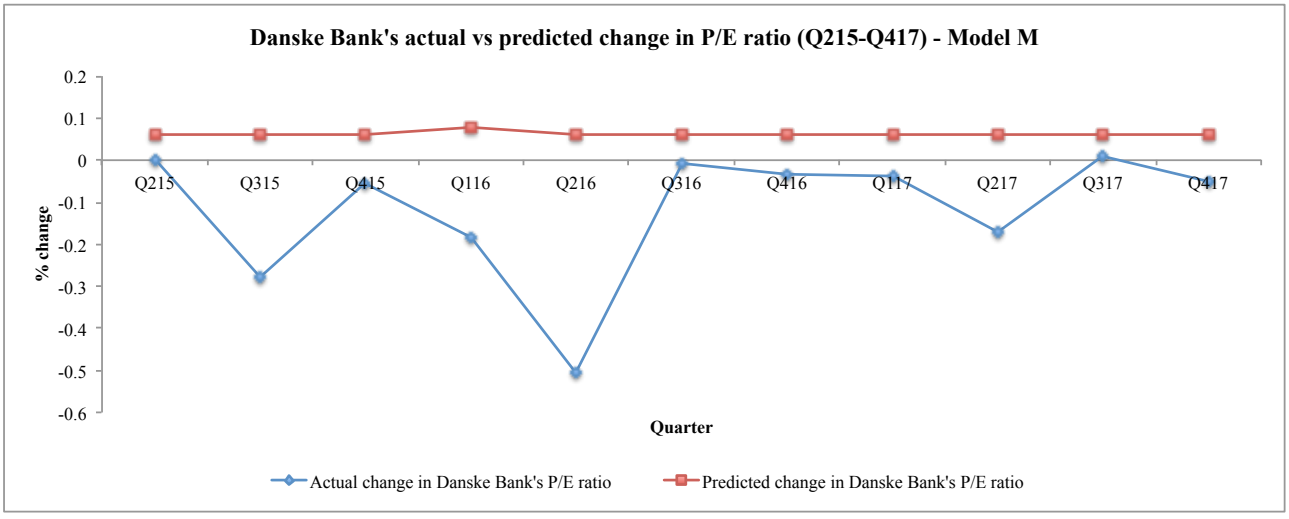

Figure 37: Model M: Danske Bank's actual vs. predicted change in P/E ratio.

- *Bavarian Nordic*

This section identifies one significant correlation between the change in Bavarian Nordic's P/E ratio in one quarter and the change in the total sum of the negative sentiments about Bavarian Nordic the quarter before. Derived from Figure **25**, the mathematical expression is:

(L) $$f(x_0) = 0.041x_0 + 0.030$$

The slope of the regression line is positive (Figure **38**). This suggests that negative sentiments published on Twitter are correlated to an increase in Bavarian Nordic's P/E ratio. If this reflects a true cause and effect, and given that an increase in the P/E reflects an increase in stock price, it could suggest that negative sentiments on Twitter induce stockowners to buy more stocks. This could suggest that investors perceive the stock price as being undervalued.

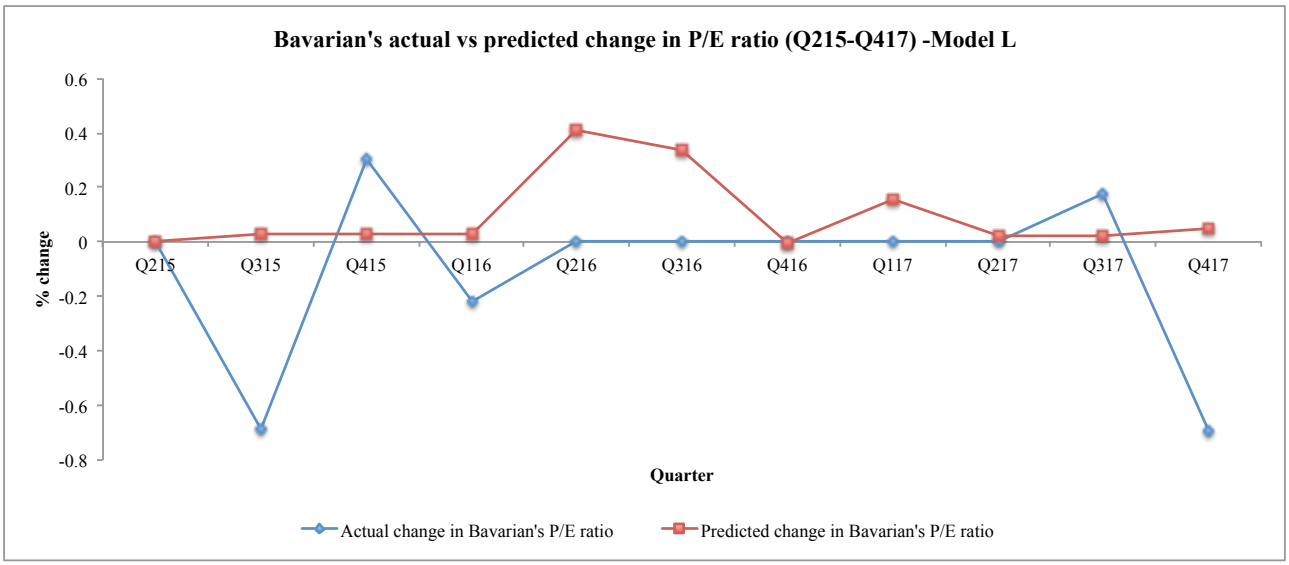

Figure 38: Model L: Bavarian Nordic's actual vs. predicted change in P/E ratio.

8.5. *Sub-conclusion on research question 3*

In this part of the study a significant correlation between Genmab's social media and stock performance is found in nine instances. P/E turns out to be the relevant parameter in seven out of the significant correlations. The relevant social media variables include articles, portals, posts, and Twitter, while Facebook has no significant correlations to the stock performance. It should be noted that mentions published on portals influence Genmab's P/E ratio with over 22% while all other parameters has less influence. Apart from Genmab, social media activity about the four other selected Danish C20 companies are not influencing the companies' short-term stock performance within the same quarter.

8.6. *Sub-conclusion on research question 4*

A significant correlation is found between the positive mentions about Danske Bank published on portals in one quarter and Danske Bank's P/E ratio in the following quarter. Furthermore, a significant correlation between negative sentiments published on Twitter about Bavarian Nordic in one quarter and Bavarian Nordic's P/E the following quarter is found.

The findings imply that changes in the social media variables influence the development in Danske Bank's and Bavarian Nordic's P/E ratio in the following quarter. Hence, the results suggest that social media activity about Bavarian Nordic and Danske Bank's do influence the firms' long-term stock performances. However, the slopes indicate that the influence of these Social Media Data on the P/E ratios are less than 5%, and when breaking down the data to look at individual quarters, the similarity between the model and the actual data is not impressive (cf. Figure **37** and Figure **38**)

9. Discussion and limitations

Nanli et al. (2012)'s comprehensive survey indicates that the amount of research that investigates practitioners' view on the implications of sentiment analysis are limited. The methodology used in the present study allows for an investigation of the practitioners' (investors) view on the implications of using, not only sentiment analysis, but also social media in general to make decisions on whether to trade stocks. The results of the interviews and questionnaires conducted contradict Li et al. (2014)'s findings in the sense that none of the investors indicate that their trade decisions are impacted by public sentiments derived from social media. Traditional finance theories assume investors to be *rational*, but Simon (1955) questions humans' rationality when making decisions. The findings of the present study contradict traditional finance theories as nine out of the 11 significant models contain predictors that measure the polarity of the content published on social media.

In relation to the polarity of the sentiments, it is interesting that whether the sentiment is positive *or* negative seems to be indifferent for the significance level of the models. The important thing is that there are both positive and negative sentiments about the firms, and positive sentiments are not always positively correlated to the stock performance indicator, and negative sentiments not always negatively correlated to the stock performance indicator. These findings contradict Sul, Dennis and Yuan (2016) and Tetlock (2011) who state that pessimism on social media could predict down movements in stock prices.

Oliveria et al. (2013) find no correlation between social media and stock performances. When it comes to the insignificant models of this study, of which there are 1,349 (76*2 + 60*2 = 272* 5 = 1,360 − 11 = 1,349), one can conclude that it is only a few of the models that can describe a significant relationship between social media variables and stock performance indicators. Moreover, in the research paper by Tumarkin and Whitelaw (2001), it is found that posting volumes and stock prices are insignificantly correlated. This study confirms Tumarkin and Whitelaw (2001)'s findings, even though the approaches of the two studies are different. When it comes to the tweet sentiment and the message volume, Sprenger and Welpe (2010) find that these two variables can predict the stock market. The present study finds a significant relationship between the message volume about Genmab on Twitter and the number of negative sentiments posted about Genmab on Twitter. However, in the case of Genmab, tweet sentiments are not correlated to the movement of the stock performance indicator. In contrast to the Genmab-case, Bavarian's stock performance indicator is to some degree correlated to tweet sentiments, but not the posting volume as suggested by Sprenger and Welpe (2010). As an extension to Sprenger and Welpe (2010), this study finds that positive

sentiments published about Danske Bank on portals, can contribute to next quarter's movement in Danske Bank's P/E ratio.

Based on the significant models to predict next quarter's change in Bavarian's and Danske Bank's P/E ratio, this study concludes, as Sul, Dennis and Yuan (2016) also do, that marginally better returns can be generated when taking social media into account. These findings do contradict with the Efficient Market Hypothesis (EMH) and imply that market inefficiencies do exist on the Danish stock market, and that the investors buying these two stocks do behave as rationally as suggested by EMH economist. However, the findings on Genmab support the EMH as social media content published about Genmab in one quarter is significantly correlated to the stock performances in the same period. In this case, the stock performance reflects the information available about the firm on social media.

This study finds that changes in Genmab's P/E ratio and changes in Genmab's opening price are impacted by different kinds of post polarities on social media, and by publications on specific social media platforms. The timeframe that is investigated is a quarter. However, with these results, it cannot be denied if the same relationship between the same variables exists in even shorter timeframes.

The secondary data analysis of the present study is built around two blocks: the big Social Media Data and the Stock Performance Data. The limitations regarding this approach are that the study does consider any other parameters that could potentially affect Social Media Data *and* the Stock Performance Data. The limitations related to the big Social Media Data, are that this study does not test for robot created content as recommended by Waddell (2018). In addition, the present study does not subjectively assess SentiOne's sentiment classification of the posts. By doing so, the frequency of getting the right sentiment will most likely have increased, but it does also require the researcher to understand a lot of languages or spending a lot of time and effort on the translation. In addition, it would limit the amount of people that could perform the analysis, as the self-training of a sentiment classifier both require some set of skills and time. This project aims to present and conduct a study that will be understandable to a broad range of academics and practitioners.

10. Conclusion

The present study can conclude that the vast majority Danish investors' do not use social media trading. In specific, most professional investors are aware of social media trading, and some find it interesting too. However, none had a desire to use social media trading as a trading strategy. The majority of Danish private investors are not aware of social media trading. Out of those private investors who are aware of social media trading, around 10% find it interesting. Only three investors out of 98 use social media trading in practice.

The quantitative results of the present study yield 11 statistically significant correlations between social media activity and stock performance. Nine of these are related to Genmab and all correlated social media activity and stock performance within the same quarters. The two others are Danske Bank and Bavarian Nordic, and these models are able to predict the development in both firm's P/E ratio one quarter ahead based on social media variables, which may suggest that abnormal returns can be generated.

Even though the quantitative analysis yields significant results, this study is not able to conclude that there is substantial evidence for an advantage of using social media trading on the five firms: Bavarian Nordic, Coloplast, Danske Bank, Genmab and Jyske Bank. Further research is therefore necessary.

11. Recommendations for future research

For future work, this study suggests that the test of the 11 significant models is repeated within another period of time. In addition, it is recommended to investigate if the 11 significant relationships are due to a correlation *between* the dependent and independent variables *or* because the dependent variables and the independent variables are both correlated to an external parameter i.e. news as suggested by Babajide and Adetiloye (2012) and Bashir et al. (2013).

Furthermore, this study chose to focus only on five Danish C20 firms. However, it is recommended that future research examine the remaining 15 firms as well for correlation. The present study investigated Danish investors' use of social media trading. However, a bigger investigation of Danish as well as foreign traders' use of social media trading is recommended to execute. Because the present analysis indicatea a more significant volatility in stock performance of the biotech stocks (Genmab and Bavarian Nordic) it could suggest that biotech stocks may be more sensitive to social media variations.

Other suggestions on future work with small changes in the methodology compared to the methodology of the present study could be to do sentiment classifier training, instead of using SentiOne's sentiment classification. Future work could also use the index proposed by Bollen et al. (2010) or the Market Prophit index. Researchers must be aware that these two indices are trained on the basis of the S&P500, and may not necessarily work on the Danish index too. Previous studies such as Tumarkin and Whitelaw (2001) and Sprenger and Welpe (2010) looked at specific stock related content on social media and how it impacted the stock performance. Future work could repeat the structure and analysis of the present study, but just use Social Media Data published by investors, entrepreneurs and traders. Future research may also want to investigate if any non-linear models or vector auto regression analysis (i.e. Tumarkin and Whitelaw, 2001) can predict the stock market performance as this study only studies linear relationships.

12. References

- Adams, M. & Mullins, T. & Baker, R. & Thornton, B. (2007). System Design and Implementation: A Pilot Study. *SAIS 2007 Proceedings*.
- Ajzen, I. (1991). Theory of planned behaviour. *Organizational Behaviour and Human Decision Process*. Vol 50, issue 2, pp. 179-211
- Akerlof, G. A. & Schiller, R. J. (2009). How Human Psychology Drives the Economy, and Why it Matters for Global Capitalism. *Princeton University Press*.
- Alaoui, I. E. & Gahi, Y. & Messoussi, R. & Chaabi, Y. & Todoskoff, A. & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data*, Vol, 5, Issue 12.
- Ashcroft, L. & Hoey, C. (2001). PR, marketing and the Internet: implications for information professionals. *Library Management*, Vol. 22 Issue: 1/ 2, pp.68-74
- Asur, S. and Huberman, B. (2010). Predicting the future with social media. *IEEE/WIC/ACM International Conferences, Web Intelligence and Intelligent Agent Technology*. Available online: http://dx.doi.org/10.1109/wi-iat.2010.63
- Babajide, A. & Adetiloye, K. A (2012). Investors' Behavioral Biases and the Security Market: An Empirical Study of the Nigerian Security Market. *Accounting and Finance Research,* Vol. 1, No. 1
- Babbie, E. (2010) The practice of social research. *12th Edition, Wadsworth, Belmont.*
- Bagnoli, M. & Beneish, M. D. & Watts, S. G. (1999). Whisper Forecasts of Quarterly Earnings Per Share. *Journal of Accounting & Economics,* Vol 28, No. 1.
- Bakar, S. & Yi, A. N. C. (2016). The Impact of Psychological Factors on Investors' Decision Making in Malaysian Stock Market: A Case of Klang Valley and Pahang. *Procedia Economics and Finance,* Vol. 35, p. 319-328
- Banerjee, A. & Chitnes, U. B. & Jadhav, S. L. & Bhawalkar, J. S. & Chaudhury, S. (2009). Hypothesis testing, type I and type II errors. *Industrial Psychiatry Journal,* Vol. 18, Issue 2, pp. 127-131.
- Barber, B. M. & Lyon, J. D. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics*, Vol 43, Issue 3, pp. 341-372
- Barnett, V & Lewis, T. (1995), Outliers in Statistical Data. *Chichester: John Wiley and Sons.* 269, p. 7
- Barry, T. E. & Howard, D. J. (1990). A Review and Critique of The Hierarchy of Effects in Advertising. *International Journal of Advertising*, 9 (2), 98-111
- Bashir, N. Y. & Lockwood, P. & Chasteen, A. & Noyes, I. & Nadolny, D. (2013). The ironic impact of activists: Negative stereotypes reduce social change influence. *European Journal of Social Psychology* 43(7).
- Beckman, T. J & Cook, D. A. (2007). Developing scholarly projects in education: A primer for medical teachers. *Medical Teacher,* Issue 29, 210-218.
- Beigi, G. & Hu, X. & Maciejewski, R. & Liu, H. (2016). *An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief*, viewed on 2 April, 2018. <http://www.public.asu.edu/~gbeigi/files/BeigiSentimentChapter.pdf>
- Berger, H. & Woitek, U. (2001). Does Conservatism Matter? A Time Series Approach to Central Banking. *CESifo Working Paper Series*, No. 190.
- Bikas, E. & Jureviciene, D. & Dubinskas, P. & Novickyte, L. (2013). Behavioral Finance: The Emergence and Development Trends. Procedia - Social and Behavioral Sciences, vol 82, p. 870-876
- Bissattini, C. & Christodoulou, K. (2013). Web Sentiment Analysis for Revealing Public Opinions, Trends and Making Good Financial Decisions. Available at SSRN: https://ssrn.com/abstract=2309375 or http://dx.doi.org/10.2139/ssrn.2309375
- Bland, M. (2000). An Introduction to Medical Statistics. *Oxford Medical Publications,* 3rd edition.
- Bloomberg (2018). *In One Tweet, Kylie Jenner Wiped Out $1.3 Billion of Snap's Market Value,* viewed 10 April 2018. <https://www.bloomberg.com/news/articles/2018-02-22/snap-royalty-kylie-jenner-erased-a-billion-dollars-in-one-tweet>
- Bollen, J. & Mao, H. & Zeng, X. (2010). Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8

- Bordino, I. & Battiston, S. & Caldarelli, G. & Cristelli, M. & Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PloS One*, Vol 7, No. 7, pp.1-17. Available online: http://dx.doi.org/10.1371/journal.pone.0040014
- Boudoukh, J. & Richardson, M. (1993). Stock Returns and Inflation: A Long-Horizon Perspective. *The American Economic Review*, pp. 1346-1355
- Boykin, D. F. (2017). Price Prediction: Determining Changes in Stock Pricing through Sentiment Analysis of Online Consumer Reviews. *UNLV Theses, Dissertations, Professional Papers, and Capstones*
- Bryman, Alan (2010). The Research Question in Social Research: What is its Role? *International Journal of Social Research Methodology*, Vol 10, pp. 5-20.
- Bureau (2017). Social Medier 2017 Statistik i Danmark, Viewed 29 March https://bureau.dk/sociale-medier-2017-statistik-danmark/
- Burn-Murdoch, J. (2013). *Social media analytics: are we nearly there yet?* Viewe 18 April, 2018. < https://www.theguardian.com/news/datablog/2013/jun/10/social-media-analytics-sentiment-analysis>
- Business Insider (2018). *24 mind-blowing facts about Warren Buffett and his $84.7 billion fortune*, viewed on 20 April, 2018. < http://www.businessinsider.com/facts-about-warren-buffett-2016-12>
- Børsen (2018). *C20*, viewed 19 January, 2018. <http://borsen.dk/kurser/danske_aktier/c20.html>
- Caroe, P. (2016). *What is the purpose of investing?* Viewed on 26 March, 2018 <http://financeinnovationlab.org/purpose-of-investing/>
- Chaffey, D. (2017). *Social media listening tool comparison*, viewed on 4 May, 2018. <https://www.smartinsights.com/social-media-marketing/social-media-analytics/social-media-listening-tool-comparison/>
- Choi, H. & Varian, H. (2012). Predicting the present with Google trends. *Economic Record, Special Issue: Selected Papers from the 40th Australian Conference of Economists*, 88(1), pp.2-9.
- Chyan, A. & Hsieh, T. & Lengerich, C. (2011). A Stock-Purchasing Agent from Sentiment Analysis of Twitter. Available at: http://cs229.stanford.edu/proj2011/ChyanHsiehLengerich-A_Stock-Purchasing_Agent_from_Sentiment_Analysis_of_Twitter.pdf
- CNBC (2016). *Donald Trump just took a shot at Boeing in Trump Tower*, viewed 9 April 2018. <http://www.cnbc.com/2016/12/06/boeing-shares-slide-after-trump-says-air-force-ones-cost-out-of-control.html>
- CNN (2018). *Snapchat stock loses $1.3 billion after Kylie Jenner tweet*, viewed 10 April 2018. <http://money.cnn.com/2018/02/22/technology/snapchat-update-kylie-jenner/index.html>
- Cohen, J. (1998). Statistical Power Analysis for the Behavioral Sciences, *Lawrence Erlbaum Associates.*
- Cooke, M. & Buckley, N. (2008). Web 2.0 social networks and the future of market research. *International Journal of Market Research*, Vol. 50, no. 2, p. 267-292
- Corea, F. & Cervellati, E. M. (2015). The power of micro-blogging: how to use Twitter for predicting the stock market. *Eurasian Journal of Economics and Finance*, 3(4), 2015, 1-7
- Cross, F. (1973). The Behavior of Stock Prices on Fridays and Mondays. *Financial Analysts Journal.* Available online: hhttps://doi.org/10.2469/faj.v29.n6.67
- Crunchbase (2018). *SentiOne*, viewed on 29 March 2018. <https://www.crunchbase.com/organization/sentione>
- Dahiru (2008): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111019/#R4
- Darlington, R. B. & Hayes, A. F. (2017). Regression Analysis and Linear Regression Models. *The Guilford Press.*
- De Bondt, W. F. M. (1998). A portrait of the individual investor. *European Economic Review*, Vol 42, Issues 3-5, pp. 931-844
- Deadline (2018). *Amazon Stock Pummeled Monday Following Trump Twitter Rant*, viewed 9 April 2018. <http://deadline.com/2018/04/amazon-stock-pummeled-monday-trump-twitter-rant-1202356984/>
- Dickerson, K. & Min, Y.I. & Meinert, C.L. (1992). Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. Journ Amer Med Assoc. 1992;263:374–378. [PubMed]

- DMR (2018). How Many People Use Facebook, Youtube, Twitter and Other Social Media (2018), Viewed 1 April, 2018 https://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/
- HD (2018). Telephone Interview. *Interview with HD from Carnegie* (see Appendix 2C).
- Druva (2009). Understanding Data Deduplication, viewed 8 April, 2018. <https://www.druva.com/blog/understanding-data-deduplication/>
- East, R. (1993). Investment decisions and the theory of planned behaviour. *Journal of Economic Psychology*. Vol. 14, Issue 2, pp. 337-375
- EUI (2018). *Datastream (Thomson Reuters)*, viewed 27 March 2018. <https://www.eui.eu/Research/Library/ResearchGuides/Economics/Statistics/DataPortal/datastream>
- Evangelopoulos, N. & Magro, M. J. & Sidorova, A. (2012). The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices? *The International Journal of an Emerging Transdiscipline,* Vol. 15, 247-268
- Field, Andy (2013). *Discovering statistics using SPSS*. London: SAGE.
- Forbes (2018). *Warren Buffett,* viewed on 20 April, 2018. https://www.forbes.com/profile/warren-buffett/
- Fortune (2017). *United Airlines Stock Drops $1.4 Billion After Passanger-Removal Controversy*, viewed 9 April 2018. <http://fortune.com/2017/04/11/united-airlines-stock-drop/>
- French, K. R. (1980). Stock returns and the weekend effect. *Journal of Financial Economics*, Vol 8, Issue 1, pp. 55-69.
- Fun, L. P. & Basana, S. R (2012): Price Earnings Ratio and Stock Return Analysis. *Jurnal Manajemen Dan Kewirausahaan,* Vol. 14, No. 1, pp. 7-12.
- Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management* 35 (2015) 137–144
- Garcia, L. P. F. & Carvalho, A. C. P. L. F. & Lorena, A. C. (2013). Noisy Data Set Identification. *International Conference on Hybrid Artificial Intelligence Systems*, pp. 629-638
- Gibbons, M. R. & Hess, P. (1981). Day of the Week Effect and Asset Returns. *The Journal of Business,* Vol. 54, issue 4, p. 579-596
- Gibson, S. (2016). Writing the Theoretical Framework Chapter , viewed 11 April 2018 <http://ccms.ukzn.ac.za/Libraries/staff-documents/Theoretical_Framework_handout.sflb.ashx>
- Goldberger, A. (1991). Multicollinearity. A Course in Econometrics. *Harvard University Press,* pp. 245-253
- Gutman Library at Harvard Business School. *Literature Review,* viewed 4 April 2018, <https://guides.library.harvard.edu/literaturereview>
- Hanna, R. & Rohma, A. & Crittendenb, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons*, Vol. 54, Issue 3, pp. 265-273.
- Hassan, S. & Nadzim, S. Z. A. & Shiratuddin, N. (2015). Strategic Use of Social Media for Small Business Based on the AIDA Model. *Procedia – Social and Behavioral Sciences,* Vol 172, pp. 262-269
- Haughton, D. (2013). Shaping the Future of Business Education. *Palgrave Macmillan*, p. 94-106
- Heale, R. & Twycross, A. (2015). Validity and reliability in quantitative research. *Evidence-Based Nursing,* Vol. 18, Issue 3, pp. 66-67
- Heer, J. & Segel, E. (2010). Narrative Visulalization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics*, Vol 16, Issue 6.
- Hirschleifer, D. (2007). Investment Theory. *Journal of Financial Economics,* Vol. 81, No. 2
- Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*. https://corescholar.libraries.wright.edu/cse/161
- Hoffman, D. L. & Fodor, M. (2010). Can you measure the ROI of your social media marketing? *Sloan Management Review* 52 (1), 41.
- AV (2018). Telephone Interview. *Interview with AV from Danske Bank* (see Appendix 2A).
- Infront (2018). *What is the difference between a private investor and a professional investor?* Viewed 7 March, 2018 http://infrontfinance.com/support-downloads/faq/what-is-the-difference-between-a-private-and-professional-investor/

- Investiopedia (2018). *Homoskedastic*, viewed on 16 April, 2018. <https://www.investopedia.com/terms/h/homoskedastic.asp>
- J. P. Morgan (2017). *Big Data and AI Strategies*, viewed on 22 March, 2018. Online availability: <http://valuesimplex.com/articles/JPM.pdf>
- PJ. (2018). Telephone Interview. *Interview with PJ from Nordea* (see Appendix 2B).
- John R. Crawford and Julie D. Henry (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, Vol 43, pp. 245–265
- Jussila, J. J. & Kärkkäinen, H. & Aramo-Immonen, H. (2013). Social media utilization in business-to-business relationship of technology firms. *Computers in Human Behavior*, vol. 30, 606-613
- Keim, D. B. & Stambaugh, R. F. (1984). A Further Investigation of the Weekend Effect in Stock Returns. *The Journal of Finance*, Vol. 39, No. 3, p. 819-835
- Keiser Education (2017). *11.3 Steps Used in Hypothesis Test*s, viewed 11 April, 2018. <https://onlinecourses.science.psu.edu/stat100/node/64>
- Kengatharan, L. (2014). The Influence of Behavioral Factors in Making Investment Decisions and Performance: Study on Investors of Colombo Stock Exchange, Sri Lanka. *Aisan Journal of Finance and Accounting*, Vol, 6, No. 1
- Kennedy, H. (2012). Perspectives on sentiment analysis. *Journal of Broadcasting & Electronic Media*.
- Kessler, S. (2014). *The Problem With Sentiment Analysis*, viewed on 3 April, 2018 <https://www.fastcompany.com/3037915/the-problem-with-sentiment-analysis>
- Kooijman, J. F. (2014). Stock market prediction using social media data and finding the covariance of the LASSO. Viewed on 12 April, 2018. <https://repository.tudelft.nl/islandora/object/uuid:588ea23b-4723-4332-bc9f-4b7aec8f8b66/datastream/OBJ/download>
- Lakonoshok, J. & Shleifer, A. & Vishny, W. (1992). The impact of institutional trading on stock prices. *Journal of Financial Economics*, Vol. 32, Issue 1, p. 23-43
- Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc., p. 949.
- Lassen, N. B. & Madsen, R. & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*
- Li, Q. & Wang, T. & Li, P. & Liu, L. & Gong, Q. & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Science*, 278, 826-840
- Liddy, E.D. (2001). Natural Language Processing. *Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc.
- Logunov, A. and Panchenko, V. (2011). A Tweet in Time: Can Twitter Sentiment analysis improve economic indicator estimation and predict market returns? Available at: https://www.business.unsw.edu.au/About-Site/Schools-Site/Economics-Site/Documents/A-LOGUNOV_A_Tweet_In_Time.pdf
- M. F. Rahman. (2011). *Reform the regulator*, viewed 20 April, 2018 <http://www.thedailystar.net>
- Malkiel, B. G. (1989). Efficient Market Hypothesis. *Eatwell, J. Milgate M. Newman P. (eds) Finance. The New Palgrave. Palgrave Macmillan, London*
- Market Prophit (2015). *Market Prophit Launches First Social Media Sentiment Stock Market Index*, viewed on 24 March, 2018. <https://www.prnewswire.com/news-releases/market-prophit-launches-first-social-media-sentiment-stock-market-index-300081778.html>
- MarketSmith (2018). *Stock Market Indicators*, viewed 20 April, 2018. <https://marketsmith.investors.com/stock-market/stock-market-indicators/>
- Matthews, N. L. & Orr, B. C. & Warriner, K. & DeCarlo, M. & Sørensen, M. & Laflin, J. & Smith, C. J. (2018). Exploring the Effectiveness of a Peer-Mediated Models of the PEERS Curriculum: A Pilot Randomized Control Trial. *Journal of Autism and Developmental Disorders*, pp. 1-18
- Mayer-Schönberger, V. & Cukier, K. (2013). Big Data: A Revolution That Will Transform How We Live, Work and Think. *American Journal of Epidemiology*, Volume 179, Issue 9, 1 May 2014, Pages 1143–1144.

- McGaghie, W. C., Bordage, G., & Shea, J. A. (2001). Problem statement, conceptual framework, and research question. *Academic Medicine*, *76*(9), 923-924.
- Mediaite (2017). *Toyota Stock Drops Immediately After Trump Tweet About Mexican Factory,* viewed 9 April, 2018. <https://www.mediaite.com/online/toyota-stock-drops-immediately-after-trump-tweet-about-mexican-factory/>
- Merriam-Webster (2017). *9 Financial Words With Surprising Origins*, viewed on 2 March, 2018 <https://www.merriam-webster.com/words-at-play/financial-word-origins/invest>
- Michaelson, D. & Stacks, D. W. (2011). Standardization in Public Relations Measurement and Evaluation. *Public Relations Journal Vol.* 5, No. 2
- Miles, J. (2005). *R*-Squared, Adjusted *R*-Squared. *Encyclopedia of Statistics in Behavioral Science,* Vol. 1
- Misirlis, N. & Vlochopoulou, M. (2018). Social media metrics and analytics in marketing – S3M: A mapping literature review. *International Journal of Information Management*, Vol. 38, Issue 1, pp. 270-276
- Montier, J. (2009). Value Investing. *John Wiley and Sons Ltd*, 1st Edition.
- Munzner, T. (2009) A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 15, Issue 6.
- Nanli, Z. & Ping, Z. Weiguo, L. & Meng, C. (2012). Sentiment Analysis: A literature review. *Management of Technology (SMOT)*
- Nazário, R. T. F. & Silva, J. L. E. & Sobreiro, V. A. & Kimura, H. (2017). A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance* 66, p. 115-126
- Nguyen, T. H. & Shirai, K. & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, Vol. 42, Issue 24, pp. 9603-9611
- Nisen, M. (2017). *Biotech Gets More Forgiving, But Riskier*, viewed 11 May, 2018. <https://www.bloomberg.com/gadfly/articles/2017-09-13/biotech-stocks-now-more-forgiving-but-riskier>
- O'Donoghue, T. & Punch, K. (2003). Qualitative Educational Research in Action: Doing and Reflecting (1st edition). *Routledge*.
- Odean, T. (1999). Do Investors Trade too much? *The American Economic Review,* vol. 89, no. 5.
- Oh, C. & Sheng, O.R.L. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *Proceedings from the 32nd International Conference on Information Systems (ICIS).*
- Oliveira, N., Cortez, P., & Areal, N.(2013). On the predictability of stock market behaviour using stock twits sentiment and posting volume. *Progress in Artificial Intelligence, Lecture Notes in Computer Science,* pp.355-365. Available online http://dx.doi.org/10.1007/978-3- 642-40669-0_31
- Oliver, V. (2010). 301 Smart Answers to Tough Business Etiquette Questions, *Skyhorse Publishing.*
- Online Etymology Dictionary. *Invest (v.),* viewed 01 March 2018. <https://www.etymonline.com/word/invest>
- Pang, B. & Lee, L. & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Language Processing (EMNLP*), pp. 79-86. Proceedings of the Conference on Empirical Methods in Natural Association for Computational Linguistics.
- Pellarkuri, V., Rajeswara, D. R., Lakshmi Prasanna, P. M. V. B. T. (2015). *A conceptual framework for approaching predictive modeling using multivariate regression analysis vs artificial neural network.* Journal of Theoretical and Applied Information Technology, Vol 77, number 2.
- Phillips, L. & Dowling, C. & Shaffer, K. & Hodas, N. &Volkova, S. (2017). Using social media to predict the future: A systematic literature review. Available online: <https://arxiv.org/pdf/1706.06134.pdf>
- Pineiro-Chousa, J. & Vizaino-Gonzalez, M. & Carvalho das Neves, J. (2017). Persistent voting decisions in shareholder meetings, *Psychology & Marketing*, Vol. 34, Issue 11, pp. 1050-1056.
- Rafanelli, M. (1995). Aggregate statistical data: models for their representation. *Statistics and Computing,* Vol 5, Issue 1, pp. 3-24
- Rajasekar, S. & Philominathan, P. & Chinnathambi, V. (2013). Research Methodology. <https://arxiv.org/pdf/physics/0601009.pdf>

- Ranco, G. & Aleksovski, D. & Caldarelli, G. & Grcar, M. & Mozetic, I. (2015). The Effects of Twitter Sentiment of Stock Price Returns
- Ratkiewicz, J. & Conover, M. D. & Meiss, M. & Goncalves, B. & Flammini, A. & Menczer, F. (2011). Detecting and Tracking Political Abuse in Social Media. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.* Online availability: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2850/3274>
- Ray, T. (2000). Sams teach yourself today e-trading: researching and trading stocks, bonds and mutual funds online. *Indianapolis – Sams Publishing*
- ResearchGate (2014). *What is triangulation of data in qualitative research?* Viewed on 02 April 2018. https://www.researchgate.net/post/What_is_triangulation_of_data_in_qualitative_research_Is_it_a_method_of_validating_the_information_collected_through_various_methods
- Ritter, J. R. & Chopra, N. (1989). Portfolio Rebalancing and the Turn of the Year Effect. *The Journal of Finance.* Vol, 44, issue 1.
- Rousseeuw, P. J. & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1 (1), 73-79.
- Ruiz, E.J., Hristidis, V., Castillo, C., and Gionis, A. (2012). Correlating financial time series with micro-blogging activity. *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp. 513-522.
- Ruths, D. & Pfeffer. J. (2014). Social media for large studies of behavior. *Science,* Vol. 346, pp. 1063–1064
- Ryu, D. (2012). The profitability of day trading: An emperical study using high-quality data. *Investment Analyst Journal* Vol 41, issue 75
- SAS (2018). *Natural Language Processing: What it is and why it matters*, viewed 11 April, 2018. <https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html>
- Schoen, H. & Gayo-Avello, D. & Metaxas, P. & Mustafaraj, E. & Strohmaier, M. & Gloor, P. (2013) "The power of prediction with social media", Internet Research, Vol. 23 Issue 5, pp. 528- 543.
- Science, J. of D. (2001). Invited Review: Integrating Quantitative Findings from Multiple Studies Using Mixed Model Methodology. *Journal of Dairy Science*, 84(4), pp.741–755.
- SentiOne (2018a). *Knowledge base*, viewed 29 March 2018. <https://sentione.com/knowledge/frequently-asked-questions>
- SentiOne (2018b). *SentiOne now available in 26 European countries,* viewed on 18 April, 2018 https://sentione.com/blog/sentione-social-listening-26-languages
- SentiOne (2018c). Knowledge base, viewed 8 April, 2018. https://sentione.com/glossary
- Seyhun, N. H. (1986). Insiders' profit, costs of trading, and market efficiency. *Journal of Financial Economics,* Vol. 16, Issue 2, pp. 189-212
- Shamsudin, N. & Mahmood, W. M. W. & Ismail F (2013). The Performance of Stock and the Indicators. *International Journal of Trade, Economics and Finance, Vol. 4, No. 6,* Available: <http://www.ijtef.org/papers/327-B10029.pdf>
- Sharma, A. & Branch, B. & Chgawla, C. & Qiu, L. (2013). *Explaining Market-to-Book*, viewed 22 March, 2018. <https://www.westga.edu/~bquest/2013/MarketToBook2013.pdf>
- Shen, D. & Zhang, Y. & Xiong, X. & Zhang, W. (2017). Baidu index and predictability of Chinese stock returns. *Financial Innovation,* Vol. 3, No. 4.
- SHIFT Communications (2015). *Why automated sentiment analysis is broken and how to fix it,* viewed 19 April, 2018 <http://www.shiftcomm.com/blog/why-automated-sentiment-analysis-is-broken-and-how-to-fix-it/>
- Sias, R. W. (2004). Institutional Herding. *The Review of Financial Studies,* Vol. 17, Issue 1, pp. 165-206
- Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, Vol. 69, Issue 1, pp. 99-118.
- Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind and Society*, Vol. 1, Issue 1, pp. 25-39
- Sitepronews, 2017. *SentiOne vs. Socialbakers: Which Social Listening Tool Is More User-Friendly?* Viewed on 18 April 2018 <http://www.sitepronews.com/2017/10/26/sentione-vs-socialbakers-social-listening-tool-user-friendly/>

- Sondari, M. & Subarsono, R. (2015). Using Theory of Planned Behavior in Predicting Intention to Invest : Case of Indonesia. I*nternational Academic Research Journal of Business and Technology*, Vol 1, No. 2, pp. 137-141
- Sprenger, T.O. & Welpe, I.M., 2010. Tweets and trades: The information content of stock microblogs. *Social Science Research Network Working Paper Series*, pp.1-89.
- Srihari, R. (2015). *How reliable are social analytics?* Viewed 19 April, 2018 https://econsultancy.com/blog/66466-how-reliable-are-social-analytics/
- Statista (2018). *Most popular social networks worldwide as of April 2018, ranked by number of active users (in millions)*, viewed on 1 May, 2018. <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>
- Stock, J. H. & Watson, M. W. (2003). "Caschool": Instructional Stata datasets for econometrics caschool, *Boston College Department of Economics*.
- StockCharts (2018). *Trading Strategies and Models,* viewed 22 March, 2018. <http://stockcharts.com/school/doku.php?id=chart_school:trading_strategies>
- Stokes, J. (2013). How to do Media and Cultural Studies. *2nd edition. London: Sage*.
- Sul, H. K. & Dennis, A. R. & Yuan, L. I. (2016). Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns. *Conference: Proceedings of the 2014 47th Hawaii International Conference on System Sciences*
- Tague, N. R. (2004) The Quality Toolbox, Second Edition, *ASQ Quality Press*, pp. 255–257.
- Tayal, D. & Satya, K. (2009). Comparative Analysis of the Impact of Blogging and Micro-blogging on Market Performance. *International Journal on Computer Science and Engineering*, Vol 1 (3), p. 176-182S
- Tetlock, P. (2011). All the news that's fit to reprint: do investors react to stale information. *Review of Financial Studies*, 24, (5), 1481-1512.
- The Verge (2018). *Snap stock starts plummets after Kylie Jenner declares Snapchat dead*, viewed 10 April 2018 <https://www.theverge.com/2018/2/22/17040332/snap-stock-price-kylie-jenner-tweet-snapchat-1-billion-market-loss>
- Thomas, J. J. & Cook, K. A. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics
- Thomson Reuters (2018). *Thomson Reuters Datastream*, viewed 26 March 2018. https://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/datastream-macroeconomic-analysis.html
- Trochim, W. M. K. (2006). *Reliability*, viewed on 12 March, 2018. <http://www.socialresearchmethods.net/kb/reliable.php>
- Tumarkin, R. & Whitelaw, R. (2001). New or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal,* Vol. 57, No. 3, pp. 41-51
- 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, pp.267-288.
- University of Arkansas (2018). *Literature Reviews,* viewed 24 March 2018. http://uark.libguides.com/c.php?g=78731&p=505552
- Veal, A. J. (2011). Research Methods for Leisure and Tourism. (4th edition). *Pearson Education Limited*
- FV (2018). Telephone Interview. *Interview with FV* (see Appendix 2D).
- Waddell, T. F. (2018). A Robot Wrote This? How perceived machine authorship affects news credibility. *Digital Journalism*.
- Wang, M. (2015). Literature Review of Decision Behaviour Biases of Enterprise Managers. *International Conference on Educational Technology and Economic Management.*
- Warsame, M. H. & Ireri , E. M. (2016). Does the theory of planned behaviour (TPB) matter in SUUK investments decisions? *Journal of Behavioural and Experimental Finance*, Vol 12, 93-100
- Weller, K. Accepting the challenges of social media research. *Online Information Review*, 39(3):281–289, 2015.
- Wilson, T. D. (2010). Fifty years of information behavior research. *Association for Information Science and Technology.*
- Wolfers, J. & Zitzewitz, E. (2004). Prediction Markets. *Journal of Economic Perspectives*, Vol. 18, No. 2
- Zar, J.H. (1984) *Biostatistical Analysis.* Prentice-Hall International, New Jersey, pp 43–45.

- Zhang, L.(2013). Sentiment analysis on Twitter with stock price and significant keyword correlation. *Honors Theses, The University of Texas*