Copenhagen
Business School
HANDELSHØJSKOLEN

Detecting users at risk of becoming social media addicts:

A big data approach to identification and value realization

Authors

Mark Jakobsen (93694) & Nichlas Holmgren (92275)

Supervisor

Daniel Hardt

A thesis submitted for the partial fulfilment
of the requirements for
the degree of

Master of science
In
Business Administration and Information Systems (Digitalization)

Submitted: 8 May 2019
Number of pages: 110
Character Count: 198766

Abstract

This paper present insights into whether it is possible to identify social media addicts by using big data analytical methods, and if such identification could create value for businesses and society.my.

With these questions in mind, an extensive systematic review was composed in order to explore the knowledge basis, and find current knowledge gaps.

Firstly, the current identification methods was explored and the takeaways from these discussed and taken into consideration as we started exploring big data possibilities. Our scope was changed to the identification of risk groups as research on the field, suggested that personal contact and hands-on work with a psychiatrist or other professionals is necessary for diagnosis. Pre-Scraped Twitter data was used as our data source in our work with big data. Different pre-processing methods was applied to arrive at insightful results that could be connected to existing identification methods through basic emotions. With these insights in mind, it was discussed how such identification could create value for society and businesses alike.

The findings shows that it is possible to identify social media addicts with big data analytical methods, but that this would optimally involve direct approach with users in combination to this. Furthermore, both research and real world cases proves that there is value and competitive advantage to get from the use of big data.

KeyWords: Big data, Heuristic, Social Sciences, Addicts, Natural Language Processing, Value Realization, Data management

Table of Contents

1.0 Introduction

1.1 Importance

Internet and social media has become a major part of our everyday lives. A survey by the global web index reports that people between 16-24 spends 180 minutes on average on social media each day[1], while the average user across all ages spends 142 minutes or 10% of the day. In 2012, the average time spent was 90 minutes, a 52 minutes increase across 6 years.

55% or 4,23 billion of the worlds 7,7 billion people has access to the internet, and it is estimated that social media has 2.77 billion users worldwide ("Number of social media users worldwide", 2019). This means that 65% of internet users and 36% of the population uses social media.

Research estimates that over 210 million people worldwide suffers from internet and social media addiction and that this has various negative impacts. Depression, sleep deprivation, loss of social life and isolation of oneself is just some of many outcomes caused by excessive usage of social media(Kurniasih, 2017).
Furthermore, Social Media Addiction (Henceforth abbreviated "SMA"), can lead to an array of psychological and social problems, such as a reduction in social skills and reduced relationship commitment (Abbasi, 2019).
Therefore, it is of utmost importance that society is able to identify and help these individuals. By developing a model based on former research, we believe that it is possible to identify social media addicts with data from the medias in question.
Such an identification could prove beneficial in reaching out and treating said addicts, as they do not always recognize their issues. A treatment of these people could create value, not only in their personal life, but also for society as a whole, as SMA has been proved to be a cause of poor achievements in work-related and academic environments (Kuss & Griffiths, 2011; Al-Menayes, 2015). We believe, that being able to identify, not only Social media addicts, but also key preliminary factors, is the key to rectify this emergent issue.

1.2 Motivation

The motivation for this thesis was partly fueled by a personal interest in information systems and all that follows; including, social media. The fact that we have both

[1] http://insight.globalwebindex.net/hs-fs/hub/304927/file-2615393475

experienced social media change the lives of ourselves and people around us, coupled with a massive media coverage on social media and data scandals, led us to investigate this topic.

It turned out that social media addiction has been linked to problems in today's society. These problems includes, but are not limited to: anxiety, depression, loneliness, attention deficit, hyperactivity, worse social skills, etc. The contemporary nature of the topic, and the fact that it has an effect on our own lives, nourished our desire to advance.
Our interest was only further increased when we discovered, that while there is an extensive body of work in the area, a particular identification of social media addiction lacked. Several methods such as the bergen scale (Kuss & Griffiths, 2011) has been suggested for identification, but none has been officially recognized. A limitation that they all share is the fact that direct contact with a person in question is needed. This means that for these models to work, the individual needs to recognize the issue and feel the necessity to seek treatment.
A decrease of SMA cases, could prove to have a positive impact on society and work cases, as the severe negative impacts, has fatal consequences for both personal and professional life.

This thesis will attempt to create a identification process, that instead of relying only on direct contact; identifies social media addicts with the help of big data analytics. Hopefully our work can assist in the identification and treatment of SMA cases, and hence prove a valuable tool in the battle against SMA.

1.3 Research Question

The Research questions in this thesis has partly been defined by the existing knowledge gap by using the CIMO method, explained in Table 4.1 found in a later section of the paper, and partly by the fact that the thesis in being written at Copenhagen Business School, which substantiates the second research question.

The Research Questions for this paper are as follows:
- Is it possible to identify users at risk of becoming social media addicts with big data analytical methods?
- Could such identification create value for business and society?

1.4 Word explanations

Various concepts are used throughout the paper. To make sure that the reader of the paper have an understanding of these concepts, the most common are explained below.

Addiction

"Addiction is a primary, chronic disease of brain reward, motivation, memory and related circuitry. Dysfunction in these circuits leads to characteristic biological, psychological, social and spiritual manifestations. This is reflected in an individual pathologically pursuing reward and/or relief by substance use and other behaviors. Addiction is characterized by inability to consistently abstain, impairment in behavioral control, cravings, diminished recognition of significant problems with one's behaviors and interpersonal relationships and a dysfunctional emotional responses. Like other chronic diseases, addiction often involves cycles of relapse and remission. Without treatment or engagement in recovery activities, addiction is progressive and can result in disability or premature death(American society of Addiction medicine)." (West & Brown, 2013)

Social Media

A broad term that embraces a variety of online applications and virtual communities where individuals, through the use of public profiles can create content and interact with other users.

Social Capital

"Social capital is broadly defined as "the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition" (Kuss & Griffiths, 2011)

Natural Language

"Language that has developed in the usual way as a method of communicating between people, rather than language that has been created, for example for computers:" [2]

[2] https://dictionary.cambridge.org/dictionary/english/natural-language

Bot

"A computer program that works automatically"[3] In our case, an agent that communicates almost autonomously on social media. It *"automatically generate messages, advocate ideas, act as a follower of users, and as a fake account to gain followers itself. It is estimated that 9-15% of Twitter accounts may be social bots."*[4]

1.5 Readers Guide

Introduction	Introduces the the research topic and questions the research aspire to answer providing clear objectives and the motivations and relevance for doing so. Furthermore, it contains a word explanation list.
Knowledge construction	Provides a brief view on how knowledge was constructed throughout the thesis as well as the deductive and inductive approaches chosen.
Methodology	Presents the research criteria, research design and a brief view of the data collection methods utilized.
Preliminary systematic review	A review of the literature in the field of social media addiction. The review explores current literature to create an understanding of the body of knowledge, while finding knowledge gaps in the field; creating the foundation for the further scope of the thesis.
Introducing big data	This chapter explores the use big data in the identification of SMA. It explains the data collection and processing, while summarizing the different methods utilizing the data to identify SMA risk groups. Lastly it sums up the results and connects the knowledge from the systematic review to provide insights and make recommendations for future research.
Value Realization	This chapter discusses the suggested big data method from

[3] https://dictionary.cambridge.org/dictionary/english/bot
[4] https://www.distilnetworks.com/glossary/term/social-media-bots/

	the previous chapter and how it can realize value in the specific context of rehab centers. Ultimately, it discusses the ethical considerations and social corporate responsibilities.
Conclusion	This last part of the thesis creates an overview of the findings and answers the research questions

2.0 Knowledge Construction

Both deductive and inductive reasoning is used throughout this thesis. Rasmussen et al. (2006) defines inductive reasoning as moving from *specific to general*, and deductive reasoning as moving *from general to specific*. The study is initialized with a deductive approach, followed by an inductive and ended with a deductive approach.

Due to personal interest and experience, the study began with the broad issue of social media addiction, an important and contemporary issue worth investigating. Forward, an extensive search for literature was conducted to gain information on the topic. By exploring definitions, consequences and ways of identifying SMA (systematic review), we recognized that no research had been done regarding identification of SMA through big data analytics, thus inspiring an attempt to do so. By starting with a broad issue of social media addiction, and narrowing it down to identify a specific knowledge gap, deductive reasoning was used. It was on the basis of the knowledge and issues found in the litterature, that the research questions were constructed.
However, it was quickly discovered that identification of SMA individuals was not feasible as research on the field, suggest that personal contact and hands-on work with a psychiatrist or other professionals is necessary for diagnosis. This changed the scope to the identification of SMA risk groups instead.

Next, inductive reasoning was utilized, as the thesis moved from the specific to general. Twitter data was acquired to obtain observations of users on social media. This data was then used in a combination of big data analysis techniques and social science literature, to gain insights about the identification of users at risk of becoming SMA.

Lastly deductive reasoning was used again. General conceptualizations of big data value realization and ethical considerations were applied, to discuss the possible value

realization and ethical considerations of big data to the specific context. The aim was to deduce the implications that the conceptualizations had on the findings.

Despite the simplicity of the process described above, it must be noted that it was a highly iterative process. Formulating the research questions did not happen through interpretation of the initial literature only, but was constantly revisited as interpretation of data and literature occurred.

3.0 Methodology

3.1 Research Criteria

3.1.1 Validity

Validity in a study refers to the coherency of a study. To have a valid study, there must be a logical connection from the research question to the conclusion. To evaluate this, one must look at whether the data collection method is relevant to the problem at hand, and the theoretical point of departure, by consistently asking yourself
"Are we really measuring what we think we are measuring? (Rasmussen et al, 2006)

3.1.2 Reliability

Reliability in a study refers to the regularity of the data over time, and whether the study is repeatable by other researchers. This is often also referred to as the transparency of the study, and relies on the researcher to document procedures and processes along the way. Furthermore to evaluate reliability, data collection and analytical methods are also taken into consideration (Rasmussen et al, 2006)

3.1.3 Trustworthiness

It can be argued, that the quality of a study, is measured better by the term 'trustworthiness'. The credibility of a study increases as the researcher clarifies decisions regarding the processes and procedures, thus reinforcing the trustworthiness. By doing this, the research can be seen as transparent and others who want to check whether the results of a given study are trustworthy, can do so. To fit the trustworthiness criteria, we will explain the decisions and framework of the study, both in subsequent section, but also in the study as it takes shape. Nonetheless, even though we use raw data for some

parts of the study, it has deep roots in social sciences and psychology. This means, that even by following our steps completely, we can not promise the exact same results, as it is possible that our interpretations of the phenomenon are different to others. This is due to the fact, that interpretations of social phenomenon are individually constructed, and hence subject to variance. (Rasmussen et al, 2006)

3.2 Research design

3.2.1 Systematic Review

The systematic review is a method that identifies research about the investigated topic. The design of this specific type of review involves the selection and critical evaluation of the contributions of the identified studies. This is done to analyze and synthesize the data in order to report on the results in such a way that a clear conclusion about what **is** known and what **is not** known, is facilitated. Therefore, a systematic review, is not a traditional review of literature, but a study that explores a problem, by using existing studies.

"*A systematic review has been argued to bring replicable, scientific, and transparent approach, which seeks to minimize bias and requires reviewers to summarize all existing information in a thorough and unbiased manner. More widely, systematic reviews have been argued also to have value in collating and synthesizing existing evidence across a wide range of settings.*" (Denyer and Tranfield, 2009)

3.2.2 Action Research

Action research design, is the process of systematically collecting research data on an ongoing phenomena, relative to an objective or need of that phenomena and using this data for problem-solving; a quite pragmatic approach (Akdere, 2003; Salkind, 2010; Coghlan & Brydon-Miller, 2014). By using action research design, we try to create a deeper understanding of a given situation. This is initially done by conceptualizing and particularizing the problem, to then progress with problem-solving, by creating and moving through several interventions and evaluations.

"Action research is not a library project to learn about a certain topic that interests us. It is not problem-solving in the sense of trying to find out what is wrong, but rather a quest for knowledge about how to improve" (Eileen, 2000).

No matter the definition used of *action research,* all of these have a similar process, including five phases of inquiry. First ***identifying the problem area***, second ***gather data***, third ***interpret data***, fourth ***act on evidence (plan of action)*** and lastly ***evaluate results*** (Eileen, 2000). Throughout this process, evaluation is constantly done, as it is repeated as many times as needed, to confront the data and access the effects of the intervention and whether or not an improvement has occurred. This ultimately provides a very iterative process in nature, as it is intended to cultivate a deep understanding of a given situation to conclusively improve it.
"It is an attitude of inquiry rather than a single research methodology" (Salkin, 2010).

In this case, social media addiction and whether it is possible to identify with big data analytical methods. Additionally, whether such identification can create value to business and society. The idea is to generate new knowledge in an attempt to create change for the better (Akdere, 2003; Coghlan & Brydon-Miller, 2014).

3.3 Data collection

This section will provide a short overview of the data collection process. More in-depth explanations are given in their respective sections [4] & [5].

3.3.1 Preparation

From the start, it was clear that different types of data had to be collected, as a result of the different methods used. Due to the nature of the initial systematic review, papers of both qualitative and quantitative nature were collected to fulfill the strict requirements of a proper systematic review. Furthermore to create a transparent process, it was documented in Appendix 1 & 2. Additionally, for the data analysis part of the thesis, 476 millions tweets scraped by Stanford university in 2009 was used. These tweets contains Raw data/quantitative in the form of time/date stamps and usernames, but also data of qualitative and social nature in the form of tweet content. Because of this, it was clear from the start that different approaches had to be taken with different attributes of the data.

3.3.2 Construction & Data Saturation.

For the review, we wanted to make sure, that our gathered articles would cover all imaginable aspects that could rise questions. To make sure of this the notion of data saturation was adopted.

"Failure to reach data saturation has an impact on the quality of the research conducted and hampers content validity...Students who design a qualitative research study come up against the dilemma of data saturation when interviewing study participants...In particular, students must address the question of how many interviews are enough to reach data saturation" (Fusch & Ness, 2015)

In the context of meta-analysis such as a systematic review, reaching data saturation can be problematic as the researcher makes use of already established databases to gather information. Therefore the data saturation of the research depends on the reviewed literature being saturated. This does not take away responsibility of the researcher, who still plays an important role, as one's personal lens can cause issues in recognizing data saturation. Some might believe that research can be done without bias, but as we are human beings, we will always have worldviews and therefore bias as to when the data is saturated. As data saturation is not only about quantity but also depth, a sample size believed to answer all questions, was chosen. These were reviewed to make decisions and descriptions based on the review [A1-A3]. By doing this, both rich and thick data was ensured. This is a data collection method that is systematic in its methods, and documenting the process creates transparency. This also shows that the notation of Data Triangulation (Fusch & Ness, 2015) is kept in mind by using multiple sources of data to enhance the research criterias of the study.

With 476 million tweets, the paper would be able to cover the volume amounts needed for the research. Furthermore, as tweets are restricted to contain a certain amount of information (Date/Time, GeoLocation, Content, Username, Tags) there is no way to extend on this data.

3.3.3 Execution

Out of the 60 articles and links found for the review, only 31 made the cut into the final review. Others had no new information or simply were not related to the research. The systematic review and CIMO framework, identified the knowledge gap and created the foundation for the the research questions. A way to produce answers to these questions was then sought. Based on the review, consequences, risk factors and identification methods of SMA were found and utilized to figure out, how to execute the data analysis.

By being aware of the knowledge gaps, and which factors to look for in the data, the data was pre-processed to produce different kinds of informational data structures, that later would be put into context with the systematic review, to produce insights and answers to the research questions.

3.3.4 Interpretation

The review yielded knowledge related to risk factors, consequences and identification methods. A common pattern however, was the lack of research regarding big data and SMA. Rather the research revolved around traditional Addiction and Social Media Addiction identification methods which involved contact with the user. Next, the 31 articles were summarized, and an interpretation of the information that lied within these. The purpose of the thesis is to answer following questions.

- Is it possible to identify users at risk of becoming social media addicts with big data analytical methods?
- Could such identification create value for business and society?

While we sought to answer the first question by interpreting the information gathered from the review, in unison with the information provided by the data; the question of value creation had to be engaged differently. To answer this question, litterature on the subject was used to discuss various ways to realize value from big data .

3.4 Primary Source

The encyclopedia of research design, states that the term “primary source” is used for all sources that are original, and that this term covers both published and unpublished

sources. Furthermore, a primary source might both be from the past and present, and encompass records of events that are described by first hand witnesses. This data is *"Firsthand, unmediated information that is closest to the object of study"* (Salkind, 2010). By this classification, our twitter data is a primary source as it is *"pure in the sense that no statistical operations have been performed on them and they are original"* ("Data: Types of Data" 2018). The content of the tweets is not changed from the original mediation, but can still be deliberately or unconsciously distorted due to the fact that these are social data, and the interpretations of social phenomenon are individually constructed, and hence subject to variance (Rasmussen et al, 2006).

3.5 Secondary Data

Secondary data is data that is collected from someone else, that originally collected it for another purpose, both in published and unpublished form. For example, if a researcher conducts an analysis using a survey (their primary data) and then stores it in an archive or online, for other researchers later use, it becomes secondary data. The main advantage of this data, is that it is already collected and processed, often at a very high quality, thus saving time and money. Furthermore, the availability of large data sets in today's technology influenced world, makes it an attractive option to help analyze specific issues. However, this information is mostly "impure" as statistical operations and bias may already have been applied ("Data: Types of Data", 2018).

The main disadvantage of using secondary data, is the time spent finding data that is suited to your particular research question. The challenge is therefore to find appropriate secondary data, to aid the analysis of a specific research question (Salkind, 2010).

In this case, it regards all of the articles, books and links which are used throughout the thesis. Especially in the systematic review, this type of data is represented [A1, A2].

4.0 A Systematic Review

4.1 Part Introduction

For the Systematic review, research papers about general and internet addiction on top of Social Media Addiction were included. As research states "*Online addiction is virtually indistinguishable from social media addiction with the single difference of the latter being used primarily on mobile devices*" (Al-Menayes, 2015)
Furthermore, research compares substance addiction to social media addiction, and the use of this comparison is widely adopted within the field of research.

The new digital age has brought extraordinary developments in technology which has altered the way in which many people access and use information. While this is mostly beneficial for the development of human relationships and minds, it can have detrimental behavioural implications such as social media or internet addiction.
"For a small minority of individuals, social media had a significant detrimental effect on many aspects of life including relationships, work and academic achievement" (Kuss & Griffiths, 2011)

Social media sites are virtual communities where users can create individual public profiles, interact with real-life friends, and meet other people based on shared interests.
Social media addiction is a contemporary issue which should be discussed in today's technology heavy influenced world. Social media addiction can be compared to alcohol and drug addiction, as negative consequences that follows, influences some of the most important aspects of life. Many issues related to this type of addiction are connected to psychological problems such as anxiety, depression, loneliness, attention deficit, hyperactivity and multitasking mania (Kuss & Griffiths, 2011; Cabral, 2011).

*"Some also experience adverse health consequences resulting from lack of sleep due to being online for a long time, particularly late at night" (*Al-Menayes, 2015)

SMA can also come with physical issues that contributes to the negative consequences of relationships, work and academic achievements. These issues are especially present for

certain groups such as people of younger age (Blackwell et al, 2017), and identifying these individuals is therefore of great importance.

As social media sites become normal for more people, excessive use does too. Creating an overview of methods to identify SMA is therefore helpful for both individuals as well as organisations managing SMA or alike.

This systematic review therefore aims to distinguish different tools and factors that can be used to identify social media addicts

Furthermore, the systematic review will later be used as part of a conceptual framework, to create a new method of identifying SMA using big data analysis, in order to create a method that can measure directly on social media sites.

4.2 Methodology

The CIMO Framework

This thesis aims to create an identification process, that instead of relying on direct contact; identifies social media addicts with the help of big data analytics. Moreover, the expanded body of knowledge, can hopefully contribute to the treatment of SMA cases, and thus prove a valuable tool for the struggles of SMA.

For the systematic review the CIMO (Context, Intervention, Mechanisms and Outcomes) framework will be applied. The framework draws upon the same components as the PICO framework but in a social science context (Rousseau, D. M., 2006).

Category	General Description.	Specific to the development of the research question.
Context	- Which individuals, relationships, institutional settings, or wider systems are being studied?	- Social Media Users of all ages. - The identification of key factors and tools to identify, between SMA and social media usage/data.

Intervention	- The effects of what event, action, or activity are being studied?	- The addictive qualities of social media usage
Mechanisms	- What are the mechanisms that explain the relationship between interventions and outcomes? - Under what circumstances are these mechanisms activated or not activated?	- The facilitation of addictive behaviour and SMA in individuals in a social media environment. - Understanding and exploring the facilitation, in order to create a data analytical identification method
Outcomes	- What are the effects of the intervention? - How will the outcomes be measured? - What are the intended and unintended effects?	- Better physical and mental health - Big Data Identification of Social media Addicts - Possible value in the treatment of SMA and its consequences - Intended effect would lead to proper identification and betterment, whereas the unintended effect would be that the created model proves to be unuseable

Table 4.1: CIMO Framework (Rousseau, D. M., 2014)

Through carefully reading the collected data, it is clear that there is substantial evidence to further explore the realm of social media addiction. There is a major correlation between SMA and negative consequences. Furthermore, there is no officially recognized identification method; a problem that warrants further research and method suggestions.

The research within the field, suggests that personal contact and hands-on work with a psychiatrist or other professionals is necessary for diagnosis. As we are not professionals in psychology we choose to change the scope to the identification of **People at Risk for social media addiction**, as we seek to identify risk factors off of social media data.

Taking the model and gathered knowledge into account, the research question for the paper was drafted and later finalized into the following research questions:

- Is it possible to identify users at risk of becoming social media addicts with big data analytical methods?
- Could such identification create value for business and society?

4.2.1 Process Diagram

The diagram below illustrates the steps taken to undergo the systematic review. The diagram is made to describe the process, and provide a guide to replicating the review by creating a transparent process. The diagram is created from the core principles of the systematic review
(Rousseau, D. M., 2006).

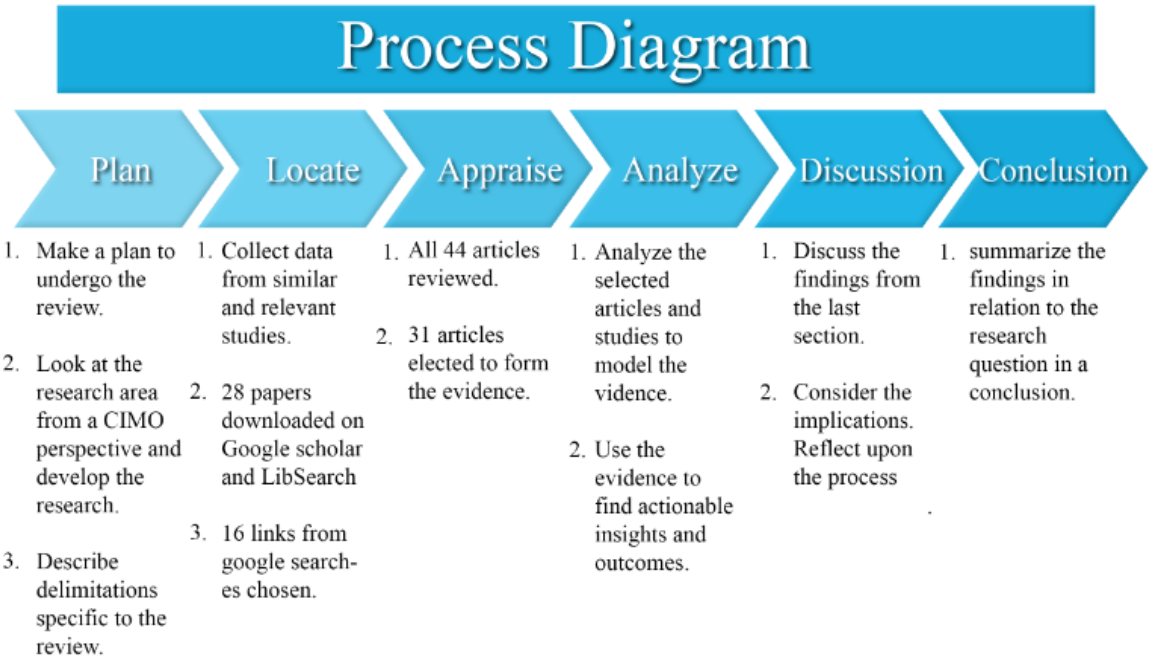


Figure 4.1: Process diagram

4.2.2 Data Collection: Keywords and Tools

This section will describe how the data for the review was collected throughout the process. First, a number of google searches were conducted, to learn if there was enough data on the topic to do a systematic review. With the results in table 1, it was established that a systematic review could be conducted on the topic. After the initial search, the scholarly literature search engine *google scholar*, was used, however acquiring access through this database to the files would be a costly affair. Subsequently CBS online library search engine *libsearch* was used, with the same search words and assumptions as previous, to find further scientific articles and evidence about the topic.

To collect a wide array of articles relating to the topic, different search combinations and sentences were used. No filters were in place for the data collection part, as we wished to examine and understand the historical development of research regarding the topic. Furthermore, it was discovered that a lot of research and articles on addiction was rather old, and since we wanted to use these to create an understanding of the underlying concepts, we decided not to make use of a timeframe filter. We chose to only download and read through articles containing at least one word of the search sentence. This was done to quickly narrow down the articles to a more manageable amount

Notice that the data in question is in regard to the systematic review and not the thesis as a whole.

Search Words	Engines	Interval	Results	Papers Downloaded	Links Chosen
"Addiction"	LibSearch, Google Scholar, Google	None	50,064 + 1,700,000 + 845,000,000	6+2	8
"Social Media Addiction"	LibSearch, Google Scholar, Google	None	23,339 + 618,000 + 365,000,000	12+3	5
"Addiction Definition"	LibSearch, Google Scholar,	None	24.875 + 758,000 + 369,000,000	1+4	3

	Google				

Table 4.2: Data Collection

4.2.3 Dataset Description

In the initial phase, 44 papers/articles were downloaded and reviewed. 13 of these were discarded and the remaining 31, selected as the primary litterature to form the evidence of the investigated topic in this systematic review. These articles all contain relevant information in regards to the topic and description of this in the CIMO framework. It's important to gather several studies, that all handle the same topic but with different aspects and perspectives (M. D. Rousseau, 2014). This is because each study has their own limitations and findings, and gathering these studies, provides more evidence for the topic while eliminating some limitations of the singular studies.

The articles all have information that is related to either addiction or social media addiction and helped build the conceptual framework. There is focus on definitions of addiction, ways of identifying SMA, consequences of SMA and other types of addiction and similarities thereof. The table in appendix [A3] consists of the name of the article, keywords, a short description of what the study is about and a description of the main findings.

4.2.4 Dataset Limitations

It is important to gather studies with a range of aspects when you research a topic. Each study will hold different limitations and a wide array of studies can overcome some of these limitations.
In this review, some of the evidence points towards general- and internet addiction, and it could be argued that gathering more evidence on the specifics of SMA could make sense. Most studies in the field suggest further research towards the topic of SMA, so while the dataset is representative of the current body of knowledge, there are signs that point towards a relatively new research field. Apart from this, most of the studies were done empirically with experiments and surveys, and not as systematic reviews or with data analytics in mind. Therefore it can prove difficult to transfer the knowledge into the

field of big data analytics; hence the specifics regarding the combinatorial use of the research in another field is not explored in this review due to the limitation.

4.2.5 Implications

The assimilation of addiction into theory and practice was in 1990 hampered by the lack of a scientifically useful definition (Goodman, 1990) in the same way the assimilation of SMA into these areas, are today. The lack of an instrument that measures social media addiction cripples the development of the research field(Eijnden et al, 2016).

A new method or concept would be of practical and theoretical value, as the current concept of SMA involves theoretical issues, that leads to considerable practical implications. Our study differs from others as our proposed method calls for a change of attention from questionnaires and other identification methods including user interaction, towards a data analysis identification method. In this method, the identification represents a new way of reaching addicts. Furthermore the proposed shift could bring implications of proactive nature, including further research into prediction models.
A hypothesis may be submitted that data representing patterns of behavioral manifestations, can be used as a proxy for observations and surveys. If so, the reviewed methods of identification can be build upon, and extended to a virtual data environment; hence be used for identification with new tools such as data analytics

We wish to validate this hypothesis by reviewing research in the field of general, internet and social media addiction, and use our findings to establish whether it is possible to identify SMA by using big data analysis.

4.3 Results

Beneath, the different topics related to SMA will be reviewed by using the corresponding literature [A3].

4.3.1 Defining Addiction

The Four keys of addiction

Addiction can imply different things to different people. According to the American Addiction Center and subsidiaries, there are four keys to addiction, and a definition linked to these that sounds as follows.

"Addiction in the repeated involvement with a substance or activity, despite the substantial harm it now causes, because that involvement was (and may continue to be) pleasurable and/or valuable"[5]

The key parts to this definition are:

1. Addiction can include both substances and/or activities

Substance addiction can include any substance that is absorbed or ingested through the body; including drugs, nicotine, coffee, improperly used prescription medications etc.
Activity addiction can include activities such as gambling, surfing the internet, watching pornography, etc. Both addictions can include substances or activities that are not required to live a full and satisfying life, but also essential things such as food and social life.

2. Addiction leads to substantial harm.

The common denominator between all definitions of addiction is that is causes substantial harm, not only to the addict, but also possibly to people around them. This phrasing is used to distinguish between occasional negative behaviour and actual addictions. Consider the difference between buying 2 or 20 packs of cigarettes a week, or the difference between spending 30 or 300 minutes on your phone a day. One might lead to a bit less money and time, the other financial burdens and loss of social life. The substantial harm can come in form of direct costs such as money, time and other tangible things, but there is also an intangible side to it. An addict can become very

[5] Definition Of Addiction. (n.d.). Retrieved from https://www.mentalhelp.net/articles/definition-of-addiction/

preoccupied with an addiction making it a central part of their life, consuming all their energy thoughts; hence, accurate appraisal of the cost/harms of addiction, is one of the defining factors.

3. Addiction is the repeated involvement of said substance/activity despite the fact that it causes substantial harm.

Any use of substance or activity can have negative consequences and lead to substantial harm, for instance, a drunk night can lead to things that should never have been said or done. It could also be that someone might use a whole day on their social media, missing an important appointment, but would we call these people addicts? One indisputable factor of addiction is the repeated behaviour despite the substantial harm. In most cases, when a behaviour or use of substance has more negative than positive consequences, people will try to limit or renounce it, but not in the case of addiction. Most people has indulged themselves in activities that might be seem as bad such as drinking, gambling, binge-watching, overeating etc. It is normal to search for pleasure, and it is known that all addictions has their beginning in the hunt for pleasure. Therefore, to distinguish between addiction and bad behaviour(even if repeated), the demonstration of loss of control is something to be considered. The underlying problem of addiction is not the enjoyment of pleasurable bad behaviour, but the inability to stop or limit this behaviour.

4. Addiction persists due a feeling of pleasure or value.

If you are on the outside of an addiction, something you might ask yourself if why someone would continue to do something that causes harm? The answers to this is simple; at the start it was pleasurable or valuable. It is important to recognize that a person can find value in a variety of situations. One person might find a form of behavior valuable as it decreases anxiety, another might find value in a temporary getaway from boredomness. Fact is, humans are genetically predisposed to repeat things we have found rewarding, it is hardwired deep in our systems and it is the reason we eat and reproduce. Due to this, only persons with prior positive experiences linked to a substance or activity are at increased risk to developing addictions. Bear in mind, that the substance or activity does not have to continue to be valuable. Over time, a lot of addictions lose their value, despite this, the addiction remains, as the new found value is the release from the cravings the initial pleasure and value caused.

These four keys can also be seen in the somewhat longer definition of addiction by the American Association of Addiction Medicine, found in section 4.2

Internet and Social Media Addiction

Research into internet addiction has developed since the mid 90's, as more cases has been detected following the boom in internet usage. The literature has several ways to describe internet addiction, mentioning concepts like

"Internet dependency," "compulsive Internet use," "problematic Internet use," "dysfunctional Internet use," and "pathological Internet use" (Al-Menayes, 2015)

Have all been used to describe what is essentially the same phenomenon 'Internet Addiction'

The term was introduced in the 1990's (Young 1999), and became generally accepted with treatment centers being created across the world. At its early stages it was a conceptualization of people who could not make a distinction between the online and offline world. Therefore, a psychiatrist or a clinician had to do a hands-on diagnose to be able to derive to the conclusion that one is an internet addict.

The Diagnostic and Statistical Manual of Mental Disorder (DSM-5), the standard classification of mental disorders used by mental health professionals in the U.S and across the world does not define Internet or Social Media Addiction as disorders. Nonetheless, in the fifth edition of the manual (DSM-5), Internet Gaming Disorder is identified as a condition that warrants more clinical research before it might be considered for inclusion as a formal disorder.

Al-Menayes states that Internet Addiction is basically indifferent from SMA, with the only difference, that the latter is primarily being used on mobile devices. While no research has dealt with the exact differences between use of mobile and computer internet/media use, it is not of impact, as the results of the addiction applies to both, as it is essentially the same medium.

The fact that most people are always online, and the popularity of free social media apps, makes it unreasonable to try and differentiate; hence, for the purpose of this paper,

there will be no differentiation between the device used to access the internet and social media.

However, we do recognize, that different social medias has different predispositions and entryways into the world of social media addiction.

"For example, the extrovert might spend much time on Facebook, repeatedly going over their profile to see the number of "likes" their latest post received. For others, with a narcissistic predisposition, Instagram may prove to be an addictive medium for them to display themselves to others with "selfies."..... significance to addictive behavior is "texting"
either directly or through social media such as Twitter and similar applications." (Al-Menayes 2015)

Furthermore, another study defines SMA as *"Being overly concerned about SNSs, to be driven by a strong motivation to log on to or use SNSs, and to devote so much time and effort to SNSs that it impairs other social activities, studies/job, interpersonal relationships, and/or psychological health and well-being"* (Andreassen and Pallesen 2014)

They also suggest that Social media Addicts often spends a lot of time thinking about social media and how they can make more time for online social networking **(salience)**. They frequently end up spending more time on social media than they intend, as they have an urge to use it more, to obtain the same amount of pleasure **(tolerance)**. Additionally, these people, make use of social media to decrease negative feelings and forget about personal issues **(mood-modification)**(Andreassen, 2015).

4.3.2 Risk Factors

Motivation and personal factors
Studies suggest that the use of social media contrasts as a function of motivation. By drawing on use of gratification theory, social media is linked to a goal-directed usage in the search for gratification and satisfaction, which bear resemblance to addiction (Kuss & Griffiths, 2011).

Additionally, research implies, that the motivation for using social media, oftentimes comes from the wish to establish and maintain different forms of social capital. This social capital can come in the form of *'Bridging'* which refers to the establishment or maintenance of weak connections between people that are solely based on knowledge sharing, instead of emotional support. This form of social capital is valuable as it offers an extensive spectrum of opportunities and means of entry to a huge knowledge base, due to the heterogeneity nature of such communities.

On the other hand, *'Bonding'* refers to strong ties, most often between family members or close friends. The motivation to use social media can come from an array of places depending on the person, but is most commonly tied to social capital and the potential increase in network size, as large numbers of weak ties(with possible development intro strong ties) can be easily formed amongst members due to the structural characteristics of digital technology.

Earlier in the paper, we used the word 'Community' to define and characterize social media. However, it is not a community in the traditional sense. In the Social Media Community you are not seen as a member in the same way, as everyone can sign up, there is no community feeling in the 'Membership'. Furthermore, these virtual communities do not come with a feeling of shared influence or equal power allocations. Alternatively, these communities *"can be conceptualized as networked individualism, allowing the establishment of numerous self-perpetuating connections that appear advantageous for users"* (Kuss & Griffiths, 2011)

This mentioned advantage is especially beneficial for people with low self esteem, as they are known to be more comfortable with creating and maintaining social capital online rather than offline. However, this is also one of the reasons that people with low self esteem are more prone to excessive use of social media; hence, low self esteem is linked to social media addiction as a potent risk factor.

Not only people with low self-esteem are at increased risk. Any person that feels ostracized or not worthy of attention, often goes online to fill this gap in their social life and find a feeling of self-worth.

Often, people with social anxiety, use social media to find acceptance and friendship they are missing. Some clinics suggest that certain people, and especially teenagers can be at increased risk for developing SMA. This risk group includes people with anxiety and depression, as they may use social media as a way to distract themselves from negative emotions and bad thoughts. Furthermore, people with high amounts of stress may turn to destress, turning it into an addictive habit.

At increased risk, is also individuals with limited social lives or people that feels ostracized. A person with insufficient social skills is more likely to socialize online, getting a feeling that opening up to other people online is more comfortable. The overall issue with social media is the fact that it becomes a substitute for whatever social aspect that might be missing in our lives, hence leading to a feeling of value that can lead to addictive habits.

Cultural factors

FOMO

When people are anxious about relationships, it is very likely that they fear being socially excluded. Fear of missing out (FOMO) is a fear that other people are having fun without you. Everyone is on social media, posting things, sharing news and content and talking to each other, which makes it hard for you to not want to be a part of it, or rather a fear of missing out. FOMO has been linked to increased social media use and has been deemed quite useful to predict and identify SMA. FOMO consists of 10 items such as "*When i miss out on a planned get-together, it bothers me*" measured on a 5 point scale, where 1=not at all and 5 = extremely true (Blackwell, et al, 2017).

The network Effect

The idea that any network becomes more valuable as more people connect to that network. For example, because one person has a Twitter account, the other wants to as well. A study from Cornell information science further proved that this effect is difficult to manage as people have a harder time quitting social media such as Twitter or Facebook, as it has become an addiction (Elgan, 2015). This can be strongly linked to the last section of FOMO.

4.3.3 Identification of SMA

5 Signs of SMA

According to Dr. Kimberly Young, one way of identifying SMA is through the 5 signs of SMA, which illustrates how an addict feels about social media.

1. You spend a lot of time thinking about social media' or planning how to use it.

The person feels a need to use and share everything on social media. There is a focus on always engaging in online social media activities, which can often lead to regretful sharing of information, due to a clouded judgement from social media engagement (salience).

2. You feel an urge to use social media more and more.

When you don't know what to do, the only thing that is on your mind is to check for updates, new messages etc. Essentially making your default free-time activity social media engagement. This could e.g. be in between non-interactive moments at social activities (tolerance).

3. You use social media in order to forget personal problems.

An aspect of addiction is using behaviour as an escape from problems. Some therefore use social media as a distraction from personal problems (mood modification).

4. You become restless if you are prohibited from using social media

If you feel anxious or depressed when you are not connected to your network, it may be a sign of addiction, as an element of *withdrawal* appears.

5. You use social media so much it has a negative impact on your relationships.

It is more comfortable for you to interact online than offline, creating an over dependence on social media(conflict). This can ultimately sacrifice time spent on real-life social activities, which could create fearful and uncomfortable feelings associated to face-to-face communication (Dr Young. K, nd).

The Bergen Scale

The bergen scale is a six-question questionnaire[A4] anchored in addiction theory as it operationalizes social media addiction by focusing on 6 different subjects.

Salience, *mood modification, conflict, withdrawal, tolerance and relapse.* These are worded by diagnostic addiction criteria, which are answered on a five-point range from *very rarely* to *very often*. In order to grade this, polythetic scoring is used which ranges from 6 to 30, where the cut-score is set to >3 on at least four of the six questions (Andreassen et al, 2012; Andreassen, 2015; Blackwell et al, 2017).

The Intrusion Questionnaire (FIQ)

An eight-item questionnaire which measures social media dependence. The questions are based on an internet addiction scale and measures the following: *Salience, interpersonal conflict, time of use (tolerance), relapse, euphoria, loss of control, withdrawal* and *conflict with other activities.*

The response is based on a 7 scale answer from 1 = never and 7 = always. answers and where the higher the score, the larger intrusion social media has in your life. A question could e.g. be: "*I often think about social networking sites when i am not using it" (Elphinston & Noller, 2011; Pengcheng et al, 2018).*

Addictive Tendencies Scale (ATS)

A three-item questionnaire anchored in general addiction theory and research. It focuses on *salience, loss of control* and *withdrawal.*

It is scored on a seven point scale that ranges from *strongly disagree* to *strongly agree.* A higher score indicates signs of SMA (Wilson et al, 2010; Andreassen, 2015).

Social networking Website Addiction scale (SNWAS)

SNWAS is a five-item questionnaire that is based on video game addiction and addiction scales. It measures on following metrics: *time usage, withdrawal, relapse, conflict, tolerance and salience*.

These are scored on a seven-point scale ranging from *completely disagree* to *completely agree,* where a high score indicates SMA. An example of question could be: "*When i am not using this social networking website, i often feel agitated"* (Turel & Serenko, 2012; Andreassen, 2015).

Internet Addiction Criterions (ADC)

This instrument focuses on time usage and 7 symptoms of addiction to identify internet addiction. These symptoms are the following: preoccupation (strong desire for internet), withdrawal, tolerance, lack of control, continued excessive use despite knowledge of negative effects, loss of interests excluding internet and use of internet to escape or relieve a dysphoric mood.

According to these criterions, if the first two symptoms are met (preoccupation & withdrawal) along with at least one of the others, internet addiction is certain (Ran, et al, 2010).

Lastly, there have been many different ways of using and expanding these, as some have included other factors such as FOMO, attachment style, anxiety, depression and self-esteem (Blackwell, et al, 2017).
While several other methods exists, these focus on similar symptoms to the instruments mentioned; hence they are omitted.

4.3.4 Consequences

A large dependance on social media can have various consequences. By constantly engaging in addictive behaviour, despite suffering numerous negative repercussions, the consequences will eventually ensue. These are namely physical discomfort, social disapproval, financial loss or decreased self-esteem (Sussman, 2011).

Psychological problems

Similar to other addictions, the change in a person's behaviour is often associated with a relief from negative feelings or discomfort and stress (escape/control), thus creating emotional issues. A controversy with social media addicts is that they use social media, in order gain control, but instead the quite opposite occurs. Social media addicts may also use these sites to stay disconnected from their own feelings, making them unable to detach from their destructive behaviour regarding social media usage. Despite addicts describing using social media as a positive feeling, they might exchange this for 'engagement'. However, they do not feel good about anything else (activity, hobby etc.), unless it involves social media (Andreassen, 2015).

The issue lies with the addicts feeling insecure about their connections with their peers, thus prioritizing their online life instead of their real ones (Kuss & Griffiths, 2011). Ultimately this can have various psychological consequences including poor self-esteem and well being, depression, anxiety, dysfunctional coping, loneliness and attention-deficit/hyperactivity disorder (Kuss & Griffiths, 2018; Kuss & Griffiths, 2011; Andreassen, 2015). These may also be underlying issues of the individual which could have provoked the addiction, only aggravating the symptoms.

Relational problems

By preferring social media and online life to social and recreational activities, relations with real life communities suffer, as time spent with these are reduced and threatened. Essentially as SMA gradually evolves, peers stop expecting time from them, causing social withdrawnness, leaving these with a quite troubled life (Andreassen, 2015; Kuss & Griffiths, 2018).

The use of social media may also cause emotional discomfort as they isolate themselves from their surroundings, which may cause further psychological problems as stated earlier. This may negatively influence relationships at home, at work/school and socially as relationship dissatisfaction may occur from jealousy or surveillance behaviour, creating social dysfunction (Andreassen, 2015).

Furthermore this can also have negative consequence on romantic relationships as the availability of private informations on different social media profiles including status updates, comments, pictures and new friends can develop jealousy, cyberstalking (electronic surveillance), reduced relationship commitment or infidelity. Most of this happens as time is spent on online friends, rather than their significant other. This can in extreme cases lead to divorce or legal action (Kuss & Griffiths, 2011; Abbasi, 2019).

Health problems

Excessive use of social media may produce sleep difficulties. Social media addicts report more sleep problems and poorer sleep quality compared to people that are not addicted to social media. “More is better” for social media addicts, as it is in their best interest to stay online, despite this resulting in too little exercise, rest and recovery (Andreassen, 2015).

Performance problems

Social media addicts spend less time studying and working, creating a negative impact on performance as procrastination, distraction and poor time-management occur. Essentially this lowers productivity in both the workplace and educational settings, as focusing on the task at hand seems difficult for the the people addicted to social media (Kuss & Griffiths, 2018, 2018; Kuss & Griffiths, 2011).

People addicted to social media will simply spend more time focusing on their online social networks rather than foregoing the activities they are expected to in different settings, ultimately bringing a negative influence on these. Due to the digital distraction of social media, this occurs, essentially lowering efficiency, productivity and achievements(Andreassen, 2015).

The previously mentioned issues regarding, psychological problems, relational problems and health problems might also result in unfavorable influence upon efficiency and achievements. Lastly, as the internet is inseparable from today's society in almost any setting, it creates a venue for discussion on how to treat SMA properly, as total abstinence of online social networking behaviour seems impossible (Andreassen, 2015).

Lastly SMA can cause dangerous behavior, as social media use is habitual enough to result in dangerous behaviour such as checking social media while driving. However, most of the behaviour changes from excessive use of social media might rather be annoying than dangerous, yet still indicative of a societal problem which should be considered(Kuss & Griffiths, 2018).

4.3.5 Addiction Comparison

Social media addiction, despite not being acknowledged as a disorder, can include cravings that are worse than other acknowledged addictions. A study by Chicago University concluded that the cravings or urges to use social media can be much stronger than those towards cigarettes and alcohol. Social media has massive conflict with self-control as: *"Highest self-control failure rates" were recorded with media."* (Meikle 2012).

Moreover, research done by Michigan State university states that excessive use of social media and poor decision-making are highly linked, similar to excessive use of substances such as alcohol and drugs (Donnelly, 2019).

Research from Harvard University further proves that self-disclosure communication stimulates the brain much like sex and food does (Walker, 2018).

This has significant implications, as compared to other addictions, abstinence is not an option in today's society (Kuss & Griffiths, 2011). The cluster of behaviours related to heavy or excessive use of social media has, for the reasons above, become subject to much discussion and research.

4.3.6 Outcomes

The figure below shows a brief overview of the outcome from the results section, concerning the three main subjects: risk factors, identification and consequences. Especially identification is crucial for the upcoming research in this thesis, as it will later be used in unison with big data analytics, to explore whether it is possible to identify users at risk of becoming social media addicts, with big data analytical methods.

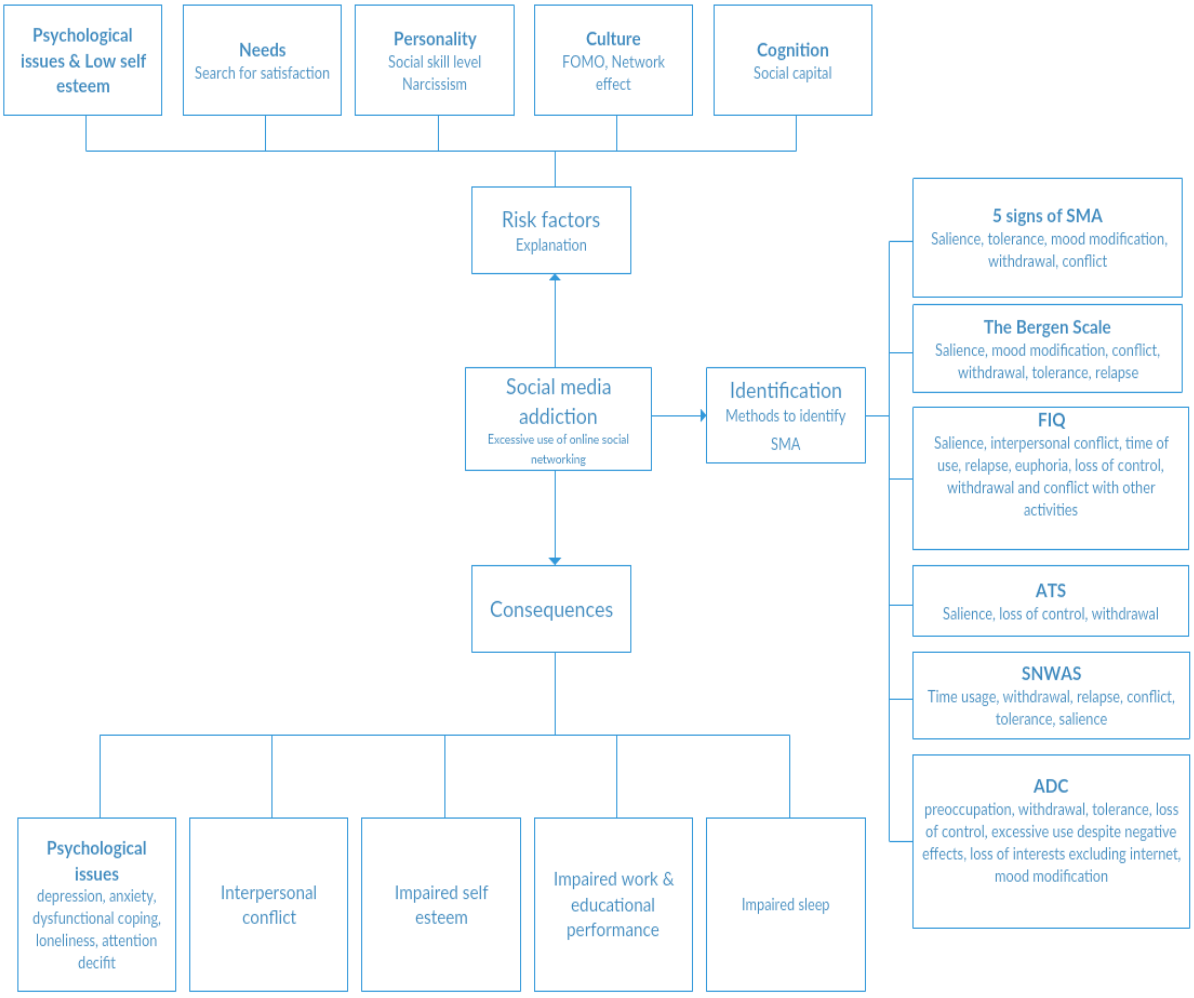


Figure 4.2: Results overview

4.4 Discussion and further implications

4.4.1 Implications for Research

There has been a lot of research associated to the topic of social media addiction. Despite this and a wide spectrum of instruments to identify SMA, there is no method that does not require direct involvement of the individuals, that may or may not be addicted to social media. Most of these use similar subjects to encounter social media addiction (salience, withdrawal, tolerance, etc.). This puts a large focus on surveys, questionnaires and interviews to solve this issue in specific cases, instead of using the data created by the media itself , to analyze the situation. Further research within this field could potentially create results aiding SMA more efficiently.

Social media is a wide term (Twitter, Facebook, Instagram etc.) and each media should be handled slightly different, due to the structure of the data. However, as previously mentioned, social media addiction and internet addiction are very similar, which makes it possible to treat these somewhat alike. Hence, for future work, it is of importance to account for the risk factors as well as the identification methods of SMA and apply these in each specific situation, especially as most of these have been proved useful in different settings (platform, demography, geography).

4.4.2 Implications for Practice

Even though there's a lot of research on the topic that points towards identification of social media addiction, it is still not officially recognized, and therefore there is no official way to identify a person as a social media addict. Furthermore, all the proposed scales requires personal contact, where our proposed method should do this with data. Even though social media is a major part of most people's everyday life, it tends to be a glorified medium. By bringing forward more research on the topic, there is a chance that we, as academics, can bring the topic into the public eye; informing people of the negative consequences. Furthermore, by identifying risk groups, we hope that treatment centers can reach out and improve on their practises by using proactive treatment methods, resulting in joint value for the identified people and numerous profit/nonprofit organisations.

4.4.3 Implications for thesis

The systematic review will prove as our knowledge base for big data analytics. The different factors in our model will all have reference in existing knowledge, and therefore have validity in our data collection and findings as our data represents true findings of the studied phenomenon. This leads to criteria and content validity for further work in the thesis, as every step regarding the data used for the review can be traced back and replicated.

4.4.4 Limitations of the Systematic Review

A systematic review is limited to existing research in the field. Hence, no new research or knowledge is created through the review. Creating a review by using former reviews can results in a meta-analysis issue, where the first review might be interpreted wrongly due to the eye of the beholder and therefore biased. Furthermore, even though, there might be a lot of research in the area, most seems to be based of surveys and interviews, which are easily affected by bias and subjectivity. Furthermore, empirical studies into the topic could prove advantageous.

5.0 Introducing Big Data

Social Media and its usage has almost tripled across the globe from 2008 to 2018 ("Number of social....", (n.d.)). This increase in adoption of social media, paired with an increase in general technology usage, has lead to increased amounts of generated and stored data. At the current rate, 2.5 quintillion bytes of data is produced every day. This means, that 90% of the worlds data has been created within the last two years. With all this data, the big data revolution is becoming progressively more relevant. The usage of big data is changing organisations, businesses, state affairs and everyday life. This immense amount of data, is a result of the evergrowing datafication; A practice that according to Schönberger & Cukier is the process of transforming several aspects of life, into data and translating it in order to create value(Schönberger & Cukier, 2013).

Even though the digitalization and datafication has assisted the progress of this revolution, the value does not lie within the revolution itself, but in the data and the potential value it brings with its variance (Madsen et al, 2016). The availability of this immense amount of data, makes it possible to obtain valuable information that would

not be obtainable with smaller data sets. Therefore, big data, enables the employment of data in new ways.

Even though the phenomena is somewhat established, an official definition does not exist. Many in the field often use the four V's of Big data adopted from the alternative definitions from Gartner(Gartner, 2016) and IBM[A5].

The four V's denotes

- **Volume** (The scale/amount of data.)
- **Velocity** (The speed at which data streams are created, stores and analyzed.)
- **Variety** (The different kinds of data, both sources and types.)
- **Veracity** (The uncertainties (noise, bias, latency, etc.) in data)

Günther et al leaves out veracity in their definition, but mentions that data is often either produced or collected for exact purposes. Therefore, depending on the variety and granularity of the data it can prove challenging to forecast which future insights that can be collected from data. They believe that the hype, publicity and ungrounded optimism concerning big data might lead organisations to assume that there is more value to be gained from the use of big data, which they end up being able to realize in practise.

The Human intelligence is of great importance when it comes to deriving and refining insights. Therefore it should be used side-by-side with algorithms, to develop organizational models that capture the value of big data (Günther et al, 2017).
Boyd, D. & Crawford, K. (2012) defines big data as a cultural, technological and academic phenomenon, which rests on the relationship between: technology, analysis and mythology. Technology refers to maximizing computation power and logarithmic accuracy to analyze data sets, while analysis focuses on utilizing the large datasets to identify patterns for economic, social, technical and legal claims. Mythology is the belief that large data sets offers a higher form of intelligence with a possibility to generate new insights in terms of general trends and patterns, which was previously impossible

Schönberger & Cukier conveys that big data is distinguished by three shifts in how we analyze data, they coin these changes; **more, messy and correlation**.

More is the capability to analyze greater amounts of data than previously, which is facilitated by the swift development in data analysis tools. In the past, the work of data analysis and research depended on subsets of data; Big data changes the way we work with data; thus moving away from data subsets and from **some** towards **all** data.

Messy refers to the cost of using all available data. Big data, is unstructured and comes in a variety of quality as it comes from an array of sources. However, it has been demonstrated that more data has an advantage over accuracy and sophistication in algorithms when it comes to, recognizing trends and gaining insights at a macro level.

Correlation is the shift from seeking causality to seeking correlation. Big data, is data-driven, and mostly, inductive in nature. Instead of a deductive hypothesis driven analysis starting from a general theory, that uses specific data to analyze and seeks to understand why. Mayer-Schönberger and Cukier argues that big data is about identifying patterns and correlations in the data, that can lead to beneficial insights and predictions. Rather than understanding why, big data is about understanding what, and this is why the sheer amount of data is advantageous.

However, a deductive approach to big data analysis is recognized by scholars and is common, for example, in the healthcare industry, where data is collected, analyzed and visualized for specific purposes(Günther et al, 2017).

Big data alters the way society works and has a large impact on different sectors; ranging from banking, insurance, science, healthcare, E-commerce and more. Not only is it a source of innovation and financial value, it also contributes great benefits to society, for instance, when used to prevent crime and recognizing/predicting suicidal ideation.

5.1 Using Big Data

This part of the paper will foremost be focusing on the technical aspects of processing data, and the results of this. Throughout the section, the steps taken all the way from choosing data to deriving insights will be explained, and in some cases shown with code snippets.

After the explanation, the results of the preprocessing will be put into context with the knowledge gained from the review; thus producing contextualized insights to answer our research questions.

Note that not all the code is shown in the paper, but for interested parties all code can be found at https://github.com/NichlasH/Thesis.

5.1.1 The Twitter API and collection methods

Twitter gives users access to different possibilities for collecting tweets from their platform.
Users can access the REST-API, from which a user can search and collect tweets from the latest 7 days, but this comes with strict limitations per user and is better used when accessing several times, rather than maintaining a connection.
Due to the scale of the project, this API is not sufficient, as a user of this service
may only collect 15 or 180 tweets every 15 minutes, depending on search criterias. The REST-API allows access to historical data if the specific tweets are known by their ID before collection, however under a limit.

The only option for instant access to historical and real-time data is Gnip; the Twitter enterprise API platform - a monetised platform, in which data can be purchased at prices from 149 - 2499$, for at most a set 5 million tweets spanning across a maximum of 30 days. Exceeding any of these thresholds will result in greater expensenses.

Finally, there is the possibility to use already scraped public datasets, available for instant download. A method that comes without economical expenses, but instead has other negative points (5.1.3).

5.1.2 The price of Data

Putting twitters premium developer pricing and Streaming-API in perspective to the data collection method used, the following table has been constructed. It includes the Total amount of collected tweets, Total data size, Used amount of data, Subscription prices, Subscription restrictions, The days it would take with a free api approach, and an Approximated price based on the assumption, that it would be possible to get a hold of the amount of premium enterprise access to the api that it would need.

Below, the Calculations used for the perspectivation are explained.

*** Approximated Price**

((Amount of Tweets/(Tweets Per Request*Amount of Requests provided with Subscription))* Price of Subscription[6]) = Price of our used Data in USD

((520000/(500*10000))*2499) = 260$

***Approximated Collection Time in Days, when using the Free, Streaming API Approach.**

"Twitter does not make public the number of connection attempts which will cause a rate limiting to occur, but there is some tolerance for testing and development."[7] Therefore, we base our calculation on a similar project (Kruuse et al, 2017), we approximate that we are able to collect tweets with an average rate of 42 Tweets/Minute.

X: Represents the uncertainty of the amount of data, as we would have been able to work with other sizes of data if we did not have to manually filter Bots and Companies out.

Y: Represents the learning curve factor due to inexperience with scraping twitter data

(Tweets*x/Tweets/Minute)/Minutes in an hour/Hours in A day = Collection time in Days

(520000*x/42)/60/24 + y= 8,7

Description	Value
Total Data Count:	467,000,000 Tweets
Total Data Size:	70,5 GB
Used Data Amount:	520000 Tweets
Approx Price*:	260$
Approx Collection Time*:	9 Days

Table 5.1: Summarized Twitter Collection Values

The table below showcases the three different approaches to collecting twitter data. If interested in a faster approach with access to historical data, the expensive premium approach can deliver that and cut the preparation time to a minimum. If interested in a free alternative, the Streaming-API can produce a restricted amount of data per api,

[6] *https://developer.twitter.com/en/pricing/search-30day*

[7] https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/connecting.html

given that one have enough computational power at hand to deal with the amount of tweets received, however this would take a long time to collect. By using data that has been scraped already, there is almost no cost and the preparation time, only consisting of finding the data. However, collecting it in this manner, grants no control of the parameters used when scraping the data, making user, hashtag and geolocation specific research and value extraction impossible.

	Pros	Cons
Streaming API	Free to collect in any amount	Very time consuminh, higher knowledge entry
Premium	Low preparation, Minimum Colleciton Time.	higher knowledge entry, Expensive
Pre-Scraped	Low preparation, No collection time, Low knowledge entry, free	No control over gathered data, can not personalize it.

Table 5.2: Pros and cons of types of data collection

A shared issue is, that if the data acquired is not fit to be used for the free- but long term-approach, there are no other ways of acquiring more data from the same period. For the costly, but fast approach, the expenses will go up with the purchased amount of tweets. By using pre-scraped data, you are not sure to find any data that fits your criteria.

5.1.3 Choice of Data For Processing

From the beginning, it was decided to use social media data for this project, as this data is at the core of the researched issue. It was however; not decided, how and where to gather this. Two approaches were considered, one being to scrape the data, allowing for more control of the extracted data. This would have allowed for the gathering of tweets with specific hashtags, or from specific users. The pros of gathering the data like this, would be two very specific corpora, as data could have been gathered from users, classified as social media addicts, instead of relying on a proxy. Furthermore, the control group could have undergone tests and interviews to be sure that they were not at risk for social media addiction, allowing handpicked data for the specific purpose.

The other choice was to download already scraped twitter data, allowing for less specific, but bigger amounts of data. As neither of the researchers had experience with scraping tweets in addition to the twitter API's restrictions on API calls for non-professional users, and high prices for less restricted access, already scraped data was chosen. Numerous datasets were explored, but ultimately a dataset scraped by stanford University was chosen. They describe the dataset as *"467 million Twitter posts from 20 million users covering a 7 month period from June 1 2009 to December 31 2009. We estimate this is about 20-30% of all public tweets published on Twitter during the particular time frame."*[8]

Selecting this dataset would also come with challenges, due to its enormous size, counting 7 files spanning from 2 to 19 gigabytes; totalling 70,5 gigabytes.

5.1.4 Natural Language Processing

Natural Language Processing(NLP) refers to the application of computational techniques in analyzing and synthesizing natural language and speech.

The human race has been communicating and writing things down for millenia. Due to thousand years of experience, our brains has evolved, and gained an extensive understanding of natural language. When we read a written piece of text, we understand the meaning of it and see it as wisdom; not data (Ackoff, 1989).

NLP is a subfield of Artificial Intelligence focused on allowing computers to process and understand natural language. The ultimate goal is to allow computers to have the same understanding of natural language as humans (Donkor, 2013). This is a feat that has proved difficult, especially as humans do not always perceive the same piece of text identically. In the case of choosing the sentiment of a text, studies shows that humans agree on between 70% and 79%, making it impossible to reach a perfect computational model(Kruuse et al, 2017). Even though Machine and Deep learning has created a shift in natural language processing with different tools, the individual processes of NLP can also help researchers process large amounts of text, leading to insights and outcomes without the use of Machine Learning, but as a joint effort between computer and man.

[8] https://snap.stanford.edu/data/twitter7.html

5.1.5 Classification and bias

Oftentimes, when you encounter a classification working with data, the data scientist will make use of a supervised or unsupervised classification method to group the data into sets, either based on rules, or patterns found in the data. In this case, two groups based on the information found in the systematic review were created. It was sought to learn whether the most mentioned factor in the review, could serve as a single-attribute proxy for the classification of the two groups. Based on Frequency data, the data was split into two groups.

A.) 500 High Frequency Users, referred to as the Risk Group.
B.) 1664 Average Frequency Users, referred to as the Control Group.

A heuristic non data-driven approach like this implements the bias of the creators and the reviews incorporated in the process. This would have been eliminated with an unsupervised approach that groups the data based on patterns found by the machine and used algorithm; instead leaving limitations based on these.

.

5.2 Methodology

5.2.1 Choice of Methods

It was deliberately chosen to import and make use of packages to create a better coding experience for ourselves. Furthermore, there is numerous forums and help to get from other users of the packages and the documentation that comes with them. The choice to implement packages decreases the complexity level, thus allowing a *learning-while-coding* process. For measuring the results, a hands-on method to measure the accuracy does not exist for this specific type of analysis. The analysis does not consist of a traditional sentiment analysis based on machine learning. Instead, due to limited experience with coding and machine learning, a Heuristic Approach in the NLP was adopted. Therefore the notion of Trustworthiness is used (See Section 3.1.3) as the criteria to prove the validity and reliability of our paper.

5.2.2 The Heuristic Approach

Heuristic refers to hand-coded functions and lack of AI/Machine Learning. The models are therefore, NOT based on a model achieved by training on a dataset, but relying on the embodiment of common-sense and domain expertise, based on the systematic

review. It is however, still iterative, and much like a learning algorithm, new approaches and methods were adopted throughout the course of the process. One example of an improvement through the iterative process is the removal of bot keywords that will be discussed in a later section. It is important to mention that it is recognized, that by adopting a heuristic approach - a sufficient way for reaching the immediate goal was chosen; knowing that this does not guarantee an optimal process or method.

The discussion of whether another approach could have proved more useful is important, especially when moving around the domain of Machine Learning with a Heuristic Approach, therefore, the limitation section of this paper, will touch upon the subject in a more thorough and detailed manner.
Furthermore, we acknowledge that this approach will not create an identification tool, but instead lead to a more comprehensive understanding of the topic, allowing us to answer the research questions at hand.

5.2.3 Word/Sentiment Analysis

For this method, a way to make use of the insights from the review was needed. For this it was decided to analyse the words from the corpora. Due to limited experience it was not possible to create a classifier that could be trained on the data, hence a simpler approach was chosen. Even though it was decided to do the analysis manually, without the use of machine learning, a way to categorize the words in the corpora was still needed. A list of positive and negative words was needed to create the structure on which the analysis would be built upon. This list would for instance classify the word “Awesome” as positive, while the word “Annoying” would be negative.

When talking about classification by labelling/tagging, there is generally two methods; hand-labelled and machine labelled. The hand-labelled data is an expensive and time consuming process, as every word is labelled by a natural language speaker (human). The machine labelled data would be labelled by a sentiment analysis program that would be trained on data, inheriting the accuracy of the machine that labelled the data, however, this less accurate process is less expensive and time-consuming, and can create large amounts of labelled data in less time.

	Pros	Cons
Hand-Labelled	Accurate data, 71-79% accuracy	Expensive to produce
Machine-Labelled	Cheap to produce	Inaccuracy inherited from machine inaccuracy

Table 5.3: Pros and cons of labelling types.

Choosing the labelling type.

For the labelled words, a free list, provided online by the University of Illinois at Chicago, was chosen; A hand-labelled opinion lexicon of 6,784 words, compiled over many years (Hu & Liu, 2014). This lexicon was later used to transform the data and store negative and positive words in separate files.

This lexicon was chosen due to the availability, and because the source has high credibility. Furthermore this decision was based upon the belief, that it would be possible to acquire hand-labelled accuracy without the cost of this; it also fit the approach, attempting to avoid the domain of machine learning.

The strength of this simple model lies in the accuracy of the provided data. The distribution of the words in the unredacted lists are shown below.

Sentiment	Count
Negative	4783
Postive	2001
Total	6784

Table 5.4:. Lexicon Sentiment Distribution

5.2.4 The DIKW Pyramid

To create a silver lining between the sections of the thesis the Data, Information, Knowledge and Wisdom (DIKW) pyramid was employed.

The DIKW pyramid is a model that has its background in knowledge management, and is used to explain the shift from data to wisdom and actionable information.

The pyramid looks at ways of extracting insights and creating value from every kind of data imaginable, in this case, big data. The Pyramid, as by its name, is illustrated as a hierarchical model in a pyramid shape. Due to this, some people call it the data-information-knowledge-wisdom hierarchy.

The DIKW pyramid is often discussed due to its structural approach. These discussions most often touch upon the exact definitions and relationships between the layers in the hierarchy. These discussions has led to changes and suggestions to extend the model with more layers. One such layer is enlightenment, which adds a dimension of truth and moral sense. While ethical considerations will be taken into account and discussed, this is not the purpose of the model in this context. Instead the model is used to define, explain and visualize the various forms that the data takes in the process, from a business perspective. We do not scrap the idea of enlightenment and knowing why, but instead adopt Jennifer Rowley's rendition of the model, referring to each layer as the actionable version of the lower layer, hence knowledge, being actionable information

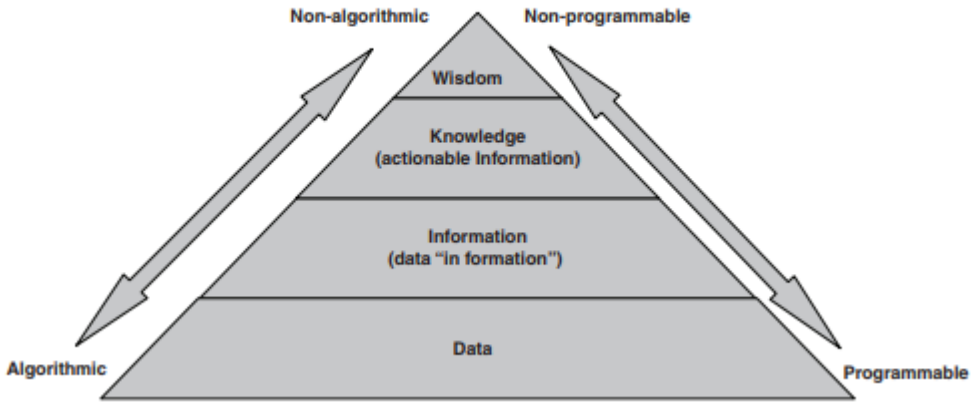


Figure. 5.1: Data, information and knowledge, Chaffey & Wood (2005).

DIKW comes with certain limitations. The model is quite linear and conveys a consistent progression of steps and stages - **information** being a contextualized advancement of **data**, as understanding increases. However, reality can be different, and knowledge can be seen as more than just the next stage of information. Nonetheless, the DIKW model is still widely used for the process of extracting value and meaning in data.

Looking at the figure, the top higher elements would be explained with the conditions of the lower element by identifying and applying a relevant transformation process. This creates an *"implicit challenge is to understand and explain how data is transformed into*

information, information is transformed into knowledge, and knowledge is transformed into wisdom." (Rowley, 2007).

There is not much sense in talking about data, information and knowledge, if these does not lead to action or outcomes. Without this, data does not represent any value, but with it, data can create value in an informed way, not only through actions, but also through knowledge turned wisdom, as seen in scientific advancements.

This paper seeks to answer if the data in question can be applied to identify SMA and thereby make decisions and gather insights To visualize this, and settle on a final model, the extended model used by AGT to convert raw, unstructured data into useful information for business decision makers, was chosen.

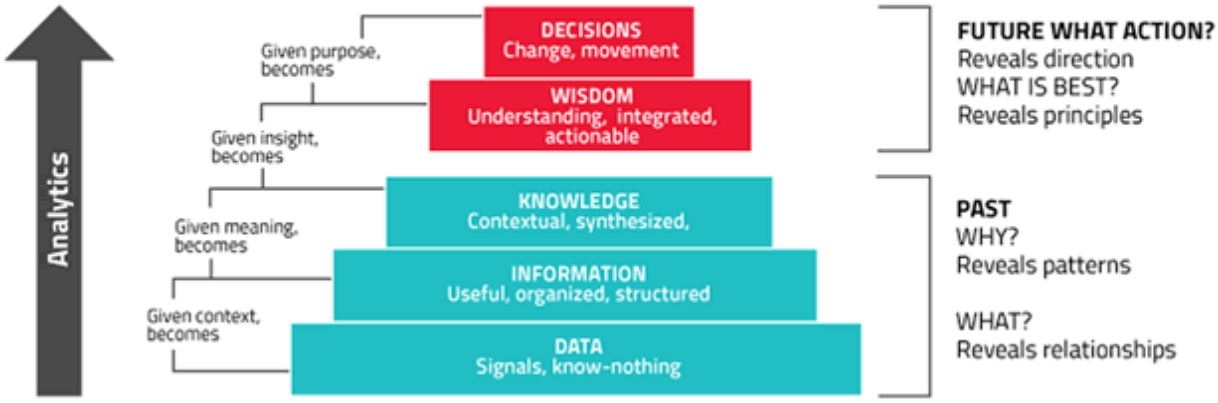


Figure 5.2:. The AGT DIKW Model[9]

5.3 Data Models

5.3.1 Bag of Words

Oftentimes, word processing tools rely on input data being represented as matrices or vectors, containing any kind and number of information; most commonly numeric value, but in this case, text.

[9] https://electronics360.globalspec.com/article/4890/optimal-analysis-algorithms-are-iot-s-big-opportunity

In a linguistics case, text data is represented as a set of all words occurring. Text can then be processed as the sum of its words is convenient. However, the obvious downside of this representation is that the sequence of words is lost - and thus the context of the word is gone. This may or may not affect the resulting analysis..

For example, the two sentences:

Life is overrated so I prefer Death instead. (5.1)
Death is but an illusion.

Would be converted to the following set:

{Life, is, overrated, so, I, prefer, Death, instead, but, an, illusion} (5.2)

Due to the data structure being a set, the word "Death" will only appear once even though it appears several times in the sentences. This set is called a bag of words, referring to its bag-like structure, containing words in no particular order.
Any sequence of words (from the bag) containing each words will produce the exact same set, resulting in each word being context free, and only counted once.
In our context, the model can be used to generate a dictionary, containing all the words in the dataset. In this case, each of the above tweets would then be described as an occurence of each word in the tweet. The tweet (5.1) would translate into

[1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0] **(5.3)**

Showing that each word occurs once. As mentioned, this leads to a loss of context.

Death is overrated so I prefer Life instead.

Would be represented the same way as in (5.3). For this reason it is difficult to get any sense of context, and the count of specific words in a text corpus, as these gets lost with the use of bag-of-words model.

5.3.2 n-gram model

One way to improve the bag-of-words model is to extend the model to group neighbouring words, creating some sort of context. If we were to pair neighbouring words in (5.1), the set would turn out:

{Life is, is overrated, overrated so, so I, I prefer, prefer Life, Life Instead} (5.4)

Two grouped elements, are called a 2-gram. A model with three neighbouring elements would likewise be called a 3-gram..

While n-grams helps preserve context, it does come at a cost. While the complexity of a bag-of-words model is linear to the amount of words in the model, a n-gram model has a space complexity of O(mn), m referring to the amount of elements in each group, in this case, 2.

Furthermore, by using n-gram models the elements are more likely to appear in new sentences than as single words, like the bag-of-words model. Furthermore, it still does not preserve the full context, and the exact count of words will get skewed; hence resulting in a tradeoff between accuracy and computational speed.

5.3.3 Tokenization

Tokenization is the process of converting a sequence of characters(Sentence/String) into a sequence of tokens (a vector with each string as an element). Tokenization is the process of separating and optionally, classifying sections of a string consisting of inputted characters. The created tokens can then be passed on to other forms of processing. For example, the text in (5.1) is not inevitably segmented with spaces as a human would do. Instead it is seen as raw input with 44 characters as seen below.

```
Sentence = "Life is overrated so I prefer Death instead."
print(len(Sentence))
44
```

Therefore, the raw input must be split into 9 tokens, given a space delimiter that matches the string. The upside of tokenization is that the user can choose their own set

of rules to identify the tokens, like specific sequences of characters, regular expressions, delimiters, dictionaries, punctuation and so on. If (5.1) is tokenized by whitespace and punctuation we would be left with following sequence.

['Life', 'is', 'overrated', 'so', 'I', 'prefer', 'Death', 'instead', '.'] (5.5)

By tokenizing, you are thus left with all the words intact, preserving context and the ability to count the amount a word occurs in our corpora, while also having the ability to split the words in a way that makes sense according to the systematic review.

5.4 Language Processing with Python and R

There is several programming languages one can use for processing natural language. Due to some experience in R it was chosen in cooperation with Python to create the scripts and processes. Python was used for text mining from the original twitter dataset, and R was used for formatting and data manipulation.

Python 3 was used to create Scripts. Furthermore, packages were used to ease the workload - of these, the most important one was **Nltk**.

Nltk(The Natural Language Toolkit), is a suite of libraries and programs that assist in symbolical and statistical NLP. Their Official website describes it as...

"A leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning."[10]

For R, mainly two packages were used, Sqldf and Dplyr. Due to SQL experience, it was an obvious choice to download the Sqldf package for R as it allowed Sql statement queries in R.

Dplyr is a data manipulation package for R that provides a consistent set of verbs to aid with the most common challenges in data manipulation, furthermore, it allowed an easy way to create randomized user subsets based on a set seed.

[10] https://www.nltk.org/

5.5 From Data To Information - Feature Engineering

In the context of DIKW, data is perceived as symbols and signs, it is raw and has no knowledge, information or apparent usefulness in its current form. Therefore, to move up the hierarchy and become useful, organized and structured information, the data must be put in context, as information is created when data is given context [Fig 5.2].

To provide context to the data, different tools and processes were used to manipulate and asses it in context to the addict identification knowledge gathered in the review. It is important to notice that the mentioned knowledge, is not knowledge gathered from the data, but exclusively the review, as the notion of knowledge in the context of DIKW is different and refers to the insight gathered from analyzing the contextualized data/information.

To turn the data into information the process of feature engineering is adopted; the process of using domain knowledge of the data to create features.

A feature is a shared attribute on which analysis or prediction can be done. The purpose of a feature is to serve as a characteristic that helps in solving a problem. Furthermore, Feature Engineering is a critical step in machine learning. While machine learning is not used in this paper, it follows the same steps in the heuristic approach, making future research and the use of machine learning possible, on these features.

5.5.1 Extracting User Frequency.

The identifications methods reviewed had several common denominators, one being time/usage. Not only was this a denominator in identification methods, but also mentioned in several definitions of Social Media Addiction. This, and the earlier stated hypothesis, led us to use the activity frequency of a user as a proxy, thus this was the first attribute to be extracted.

5.5.2 Creating a user frequency list

The first step was to create a list of users, ordered by frequency. The original data contained Author, Timestamp and Content. Below is a representation of a random picked tweet in the format of the original dataset; Author URLS are anonymized due to ethical concerns.

```
T    2009-06-25 18:45:03
U    http://twitter.com/AuthorUrl
W    @AuthorUrl but you can at least back it up with logic lol
```

To extract the frequency, we used Ubuntu for windows, as the Linux terminal grants access to bash commands such as the "Extended Global Regular Expressions print, or **egrep**; which scans a specified file line by line, returning the lines that contain a pattern matching a given regular expression. Note that by scanning line for line, no complications were caused by the size of the files.

The following expression was used:

```
egrep '^U'  Data File Path| sort | uniq -c | sort -nr > File path
to store file with counts.
```

The above expression searches the pattern of any line where U is the first character, meaning all lines with an User Url. These lines are then found and counted before they are sorted according to unique urls in descending order and written to a file.

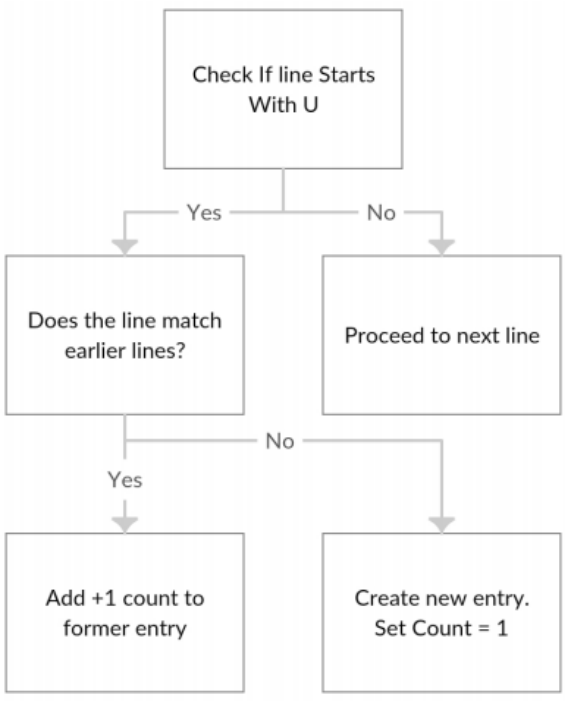


Figure 5.3: Egrep Flowchart

This command was run for all 7 files that contained twitter data, and resulted in 7 files with user frequency counts, formatted like.

```
Count U Url
Count U Url
Count U Url
```

After, the list was stripped of the redundant “U” characters. To do this, the files were opened in notepad ++, as this program is able to handle reasonably sized text files. The Replace function of notepad ++ was used, replacing “U” with ““ resulting in following format.

```
Count Url
Count Url
Count Url
```

5.5.3 Merging the lists

At this point, there was 7 lists with user frequency counts, but several users occured on different lists, so we had to merge these lists together to obtain a masterlist with all users and their summed frequencies.

```
# Merging Two lists by username
Users <- merge(L1, L2, by = "Username", all = TRUE)
# Renaming Columns to not get columns with same name
names(Users)[names(Users) == 'Frequency.x'] <- 'a'
names(Users)[names(Users) == 'Frequency.y'] <- 'b'
# Repeat for all columns
```

List Merging Code Snippet 1

This Data manipulation was done in R, by merging two lists at a time; replacing column names to avoid same name columns. After all lists had been merged into the same dataframe, there still was 8 columns, one with username, and 7 with counts.

Username	a	b	c	d	e	f	g
Url	Count	Count	Count	Count	Count	Count	Count

Table 5.5: Representation of Merged Lists

Next step was to move from this, to a list with only 2 columns, username and aggregated count. This was done by using the Sqldf plugin for R, allowing the use of Sql statements. As some values in the count columns would occur as NA, due to all users not being present in all the created lists, the Null Values has to be replaced with 0 to not run into issues when summing the frequencies.

```
# Replacing Null Values with 0
Users <- replace(Users,is.na(Users),0)
# Summing/Merging All Frequency columns under 1 column ordered by
Descending Frequency
Users = sqldf('SELECT Username, SUM(a+b+c+d+e+f+g) AS Frequency
FROM Users ORDER BY Frequency Desc')
# Writing Table to File Without Quotes and tab as delimiter
write.table(Users, file = "UserFrequencyList.txt", sep = "\t",
            row.names = FALSE, col.names = TRUE)
```

List Merging Code Snippet 2

Above code snippet changes table 5.5, to below dataframe, with 2 columns and x Rows; x being the amount of unique users.

User	Frequency
Url	Count

Table 5.6: Representation of Final frequency list

5.5.4 Grouping Users

To make comparisons within the gathered data, the data was split into two user groups. A risk group and a control Group.

The Risk Group

The initial approach to obtain a risk group was to pick a random amount of users with a frequency at least 20 times the frequency of the average active users. But after running text mining and analysis on these users, it was clear that many bots and companies hid amongst these top users; this skewed our results as, bots especially, are known for repeating the same words, leaving word counts that were non-proportional with human users. It was clear that another approach had to be taken with this issue. It was decided to make use of a time consuming process; filtering by hand. This meant, that the strength in the amount of data would be lost, as the time it would take to filter 17 million users with 200 users an hour on average would be 8500 hours, or 10 years. As this was not feasible it was decided to manually filter the top frequency users.

The criterias for the filtering process are explained with the flowchart below.

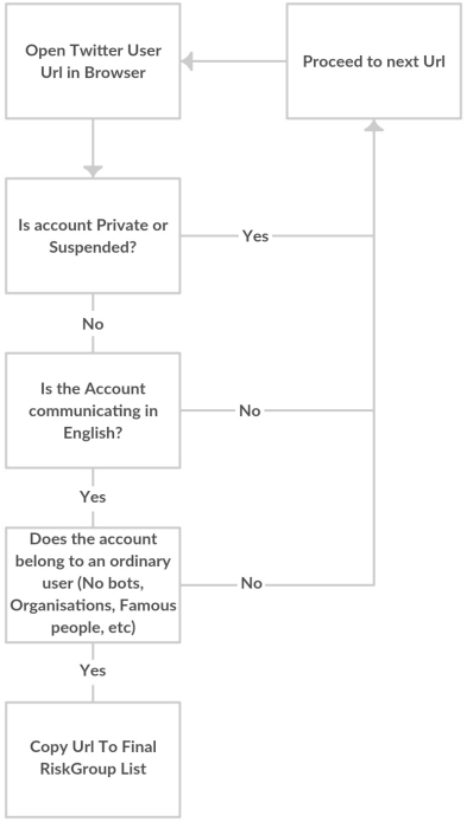


Figure 5.4: Flowchart of User Filtering for Risk Group List

This process was followed until 500 users were reached, ending up as the final list of users at risk for social media addiction; also known as, the risk group.

The Control Group

The formulation of the control group list was handled different, as it was found that there was not many bots or other problematic accounts amongst these. Furthermore, later in the process non-English words in all the text was filtered out, so non-english speaking accounts in this group did create difficulties.

For the control group, average *active* users were used; *active* being a frequency over a threshold of more than 2 activities a month, or more than 14 frequency.
Therefore, to create the control group, we filtered out any users that did not have frequency > 14.

```
# Creating List of User with more than 14 Frequency (2 Activites a month as threshold for Active User)
ActiveUsers <- sqldf("SELECT * FROM UserFrequencyList WHERE Frequency > 14")
```

Creating User Lists Code Snippet 1

Afterwards, the average frequency of the remaining users was found (210), and the Random Number Generator(RNG) aspect of the selection process was prepared by setting a seed.

```
result.mean <- mean(ActiveUsers$Frequency)
print(result.mean)
# Mean = 210
set.seed(295)
```

Creating User Lists Code Snippet 2

Then, to make sure there was enough users, while still staying in the domain of the average user, all users in a boundary of 10+/- from the average user was selected. Then, a number of users was randomly picked from the list to form the list of control users. The number of control users was found through an iterative trial and error process where the final *English* word-count was compared to the *English* word-count that the list with risk

users produced. The number of control users that resulted in the smallest discrepancy was used (1664). Finally, the Frequency column was removed and the list was written to a new file.

```
# Filtering with Average +/- 10
ControlGroup <- sqldf("SELECT * FROM UserFrequencyList WHERE Frequency >= 200 AND Frequency <= 220")
# Creating Random sample with 1664 users
ControlGroupSample <- sample_n(ControlGroup, 1664)
# Removing Frequency Column to get a list of usernames only, as frequency numbers is not used moving forward, but only to create the initial list
ControlList<- ControlGroupSample[1]
# Writing Table
write.table(ControlList, file = "ControlList.txt", sep = "\t",
            row.names = FALSE, col.names = FALSE, quote=FALSE)
```

Creating User Lists Code Snippet 3

The question might arise, why non-english accounts in the risk list was filtered out, when the possibility to filter out non-English words exists. That was a decision based on the number of users; If 500 users had been chosen, without including that criteria, it would be possible to end up with an amount of users that would create a much smaller text corpus.

5.5.5 Extracting Text

With the two groups of users done, the question of how to extract the tweets from these users was still left unanswered. Python was used to extract this as it had the Nltk library, making it easier to to perform some of the later tasks. Furthermore, dividing this process between more languages was not desirable. The paper called for two corporas, one consisting of the tweets written by the users in the risk group and another with tweets written by the control group.

At first, the big data files were loaded into python, but this approach led to Random Access Memory (RAM) problems. This made it clear, that going forward, it was

preferable to have a line by line approach; similar to the Egrep function. With no prior experience in Python the functions of the program was cut down into small steps.

First step was to open the text files consisting of the user list and tweet data.
Next, to read them line by line and check for users that are in the user group in question. If the user is found, the tweet corresponding to that user should then be printed to a new text document.

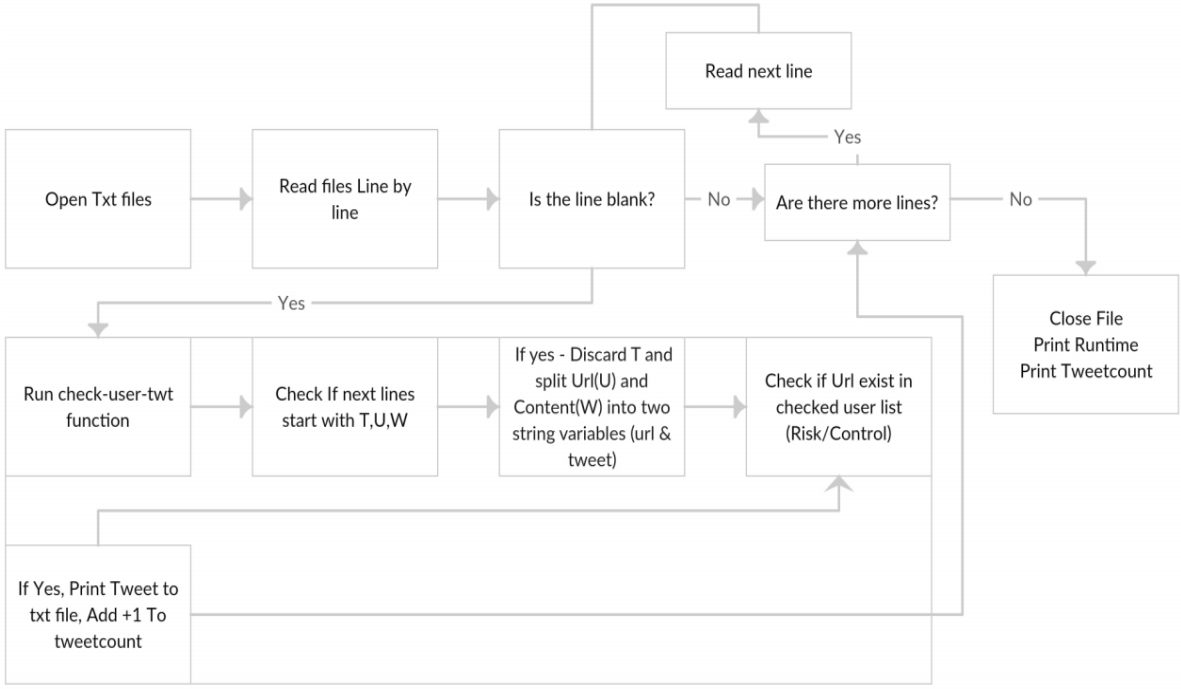


Figure 5.5: Python WriteScript Flowchart

The scripts took an Average 104 Minutes to run, and produced two corpora consisting of 269094(Control) and 249622(Risk) Tweets.

5.5.6 Tokenization

Tokenization is the task of chopping a sequence into pieces; called tokens, while simultaneously removing characters of no interest, such as punctuation. A token is often referred to as a term or word, but a distinction is necessary. In this case, a *Token is referred to* as any instance or sequence of characters that is grouped together in a semantic and useful way for later processing. A *Token Type* is a class of tokens containing the same sequence of characters. A *Term* is a (optionally normalized) *Type* with close relations to the tokens in a document such as dictionary-words in a book or

semantics in natural language. Rather than being as the exact tokens appear in a document, the term is often derived from these tokens with different normalization processes, which are discussed in sections 5.5.7 - 5.5.9. For example, if the string to be tokenized is.

"@User You're a funny person, and fuuun is a really iMportant thing.".

With the standard Nltk whitespace tokenizer, this sentence would consist of 11 tokens, but only 10 types since there are 2 instances of "a". The amount of terms will be discussed in later sections as these depend on normalization, omitting stopwords from the index, etc.

1	2	3	4	5	6	7	8	9	10	11	12
@User	You're	a	funny	person,	and	fuuun	is	a	really	iMportant	thing .

Table 5.7: Nltk Whitespace Tokenized Sentence

The biggest challenge of the tokenization phase is to correctly identify which type of tokenization to use. In this example, it is quite trivial; you chop on whitespace. This is a starting point, but there are a few tricky cases to deal with. For example, how do you deal with the use of apostrophes for possessives and contractions? Furthermore, in this case, how is casual talk, emoticons, case-preservation and user-handles handled?

As the Nltk Package for python was being used, a deeper look into the documentation of the package, revealed a special tokenize.casual.module, with a twitter aware tokenizer, designed to be flexible and easy to adapt, to new domains and tasks. It includes regular expressions found to help in handling tweets. It also has a built-in functionality to strip twitter username handles from the text and to down-case everything but emoticons, both used for the final tokenizing.

Furthermore, it has the ability to replace repeated character sequences of length 3 or greater with sequences of length 3. While the prolonged words in the corpora are

altered, a different approach was chosen, leading to more control of the process. Hence this functionality was not used.

1	2	3	4	5	6	7	8	9	10	11
you’re	a	funny	person	and	fuuun	is	a	really	important	thing

Table 5.8: Twitter aware Tokenized Sentence

5.5.7 Filtering and stopwords

Common Stopword & Punctuation

Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in nlp research, however the nltk module contains a list of 179 stop words derived from the English language.[11] This was the list that extended to include other words, that needed to be removed in the process. The first step was to extend the list with punctuation and characters such as *[(“[“,”[,”,”&”)]*. In addition to these, it was chosen to also remove common conjunctions, pronouns and re-tweet tags (“rt”) as they have no sentiment by themselves.

Taking BOTS into Account.

As the control group was not filtered manually, this group still contained some bots, made visible by the high amount of certain words. The occurrence of these words in the corpora was checked, they were mostly found as parts of repeated sentences. The bots were categorized tinto two categories

Action bots

This type of bot activates on special user actions, the most common, when a user follows or stop following another user. These sentences would consist of *“@User Stopped/Started following you”*. As the case of user handles was already considered in the tokenizing part, it was decided to deal with this type of bots by adding *([“Stopped”,” Following”])* to the stopwords.

Advertisement Bots

[11] https://github.com/NichlasH/Thesis/tree/master/Thesis%20upload/Files

These bots serves one purpose, getting users to click on links that moves them to other websites, containing the advertised product. It was discovered that most of these advertisement bots were advertising for romantic and physical pleasure services, the most common words from these bots were added to stopwords. These contains but are not limited too (["Best","Sexy","Free"]).

The Downside

Adding these words to the stopwords list comes with repercussions as these words could prove important, however, as they are removed from both corpora, the results will be comparable. It is recognized that these words can prove useful in proving salience, mood-modification etc. Nonetheless, this can also be proven with the remainder of the words in the corpora.

5.5.8 Alteration of prolonged words

As social media users generally write in a casual way, the occurrence of purposefully prolonged words, happens often. These prolonged words are used to put emphasis on words and opinions. An example could be the sentence *"My day was so goooood"*. It is reasonable to assume, that in given context, the word goooood is a prolonged version of good, and not god. The chosen alteration process makes use of the positive and negative dictionaries as follows.

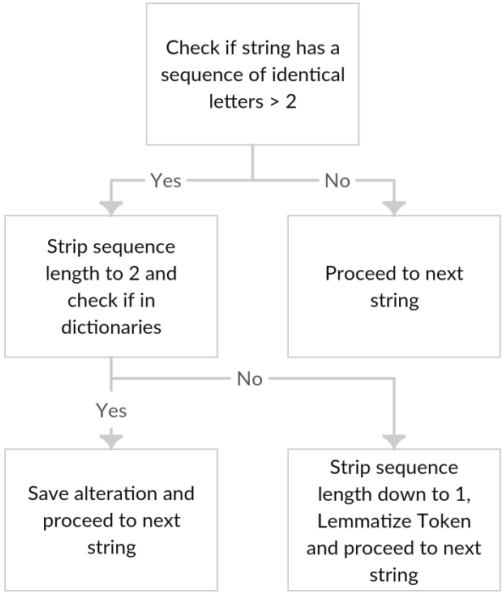


Figure 5.6: Flowchart of altering Prolonged Words

Running the above example in the flowchart, the result of the path **{Yes, Yes}**, would be

"Good". This procedure will not work in every case, as the intended word may not be contained in the dictionaries and/or read out of context. However, it does improve the corpora and provides a better count of used word.
This was written as a python function, where the word would be lemmatized after the sequence length is stripped to length 1.

```
def alter_prolonged(list):
        lemmatizer = WordNetLemmatizer()
        res = list.copy()
        for v in range(len(list)):
                i = 0
                j = -1
                w = list[v]
                while(i + 2 < len(w)):
                        if (w[i] == w[i+1] and w[i+1] == w[i+2]):
                                w = w[:i] + w[(i+1):]
                                j = i
                        else:
                                i+= 1
                if (not (w in POSITIVES or w in NEGATIVES)) and j != -1:
                        w = w[:j] + w[(j+1):]
                try:
                        res[v] = lemmatizer.lemmatize(w)
                except:
                        print("Could not lemmatize word '" + w + "'")
                        res[v] = w
                        pass
        return res
```

Alter_prolonged and Lemmatization combined python function Code Snippet

5.5.9 Lemmatization

One last step had to be taken before the pre-processing of the tweets was complete; removing word-prefixes and suffixes to reshape words to their basic forms. This makes it easier to compare and count words as they will not appear in different forms. There are two common methods to do this, *Stemming* and *Lemmatization.*

Stemming removes prefixes and suffixes in a crude heuristic process that removes ends of words to achieve this goal, which often includes removing derivational affixes[12]. It does not make use of a dictionary and suffers from the disadvantage that words has a

[12] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

chance to not be stemmed correctly; It does however benefit from faster run-times due to the nature of the process and algorithms.

The most known Algorithm for stemming is porter's algorithm[13], but as it cut words short with its ruleset and a faster runtime was not necessary, stemming was not used.

Lemmatization refers to doing things the "proper" way, with the use of a dictionary and morphological analysis of words. The goal of this process is to only remove inflectional endings; thus returning the base/dictionary form of a form; also known as the *lemma*. Furthermore, lemmatization tries to recognize whether the token is used as a verb of noun, before looking it up and returning the *lemma*.

If confronted with the token *saw*, stemming might return just *s*, whereas lemmatization would attempt to return either *see* or *saw* depending on whether the token is used as a verb or a noun. The two may also differ as stemming most commonly collapses derivationally related words, whereas lemmatization collapses the different inflectional forms of a lemma.

Examples:

1. The *Lemma* of "Better" is "good". Stemming would miss this link and return "Bett".
2. "Walking" has "Walk" as its *Lemma*. Stemming would also cut "ing" and return "Walk"
3. "Meeting" can be used as a noun or as a verb "To meet". Depending on the context, the returned *lemma* would be "Meeting" or "Meet". Stemming would return "Meet" in both cases.

The lemmatization is run after the alteration of the prolonged words, as the dictionary look-ups will then be more correct. "Betteeeeer" Would for example not return "Good" at its *lemma*. The use of lemmatization also makes it more convenient for humans to read, as stemmed sentences are not convenient to read. The ambiguity of words might cause negative effects on the analysis as the combined altering and lemmatization process will not catch all words, specially not in a case with casual language and

[13] http://snowball.tartarus.org/algorithms/porter/stemmer.html

grammar such as Twitter. A causality is caused by the nature and restrictions of the platform.

As a final Step before the results of the pre-processing could be gathered, byte-order marks(BOM) were removed, as these are useless to our analysis python marked these as tokens.

To reflect on the results, the tokenized sentence from table 5.8, ends out like this.

1	2	3	4	5	6
fun	person	fun	really	important	thing

Table 5.9: Final Processed Sentence

Even though the result is not an understandable sentence, each word is understandable by itself, and that is what is important.. Note that the words can still be found in their original context (Sentence), by searching for them in the original extracted text from section 5.5.5.

The final **ControlGroup** corpus consists of 2548578 words.
The final **RiskGroup** corpus consists of 2560102 words.

Thus, we have created two corporas with a difference of only 11524 words, making these corpora comparable in word counts as they are of similar size.

5.6 Pre-Processing Results - Information

The results of the preprocessing led to different kinds of information including which words were used most frequently in the corpora, which positive and negative sentiments the two groups made use of the most, how big a percentage of the corpora was sentiment words, etc.

5.6.1 General Token Distribution

Frequency distribution is often used in language processing. This was the first result to collect; a count of the most frequently used words in both corpora. This was done by applying the Nltk package and its included Fdist function.

A function that counts the frequency distribution of each item in the corpora. This function shows how the total number of tokens in the corpora are distributed.

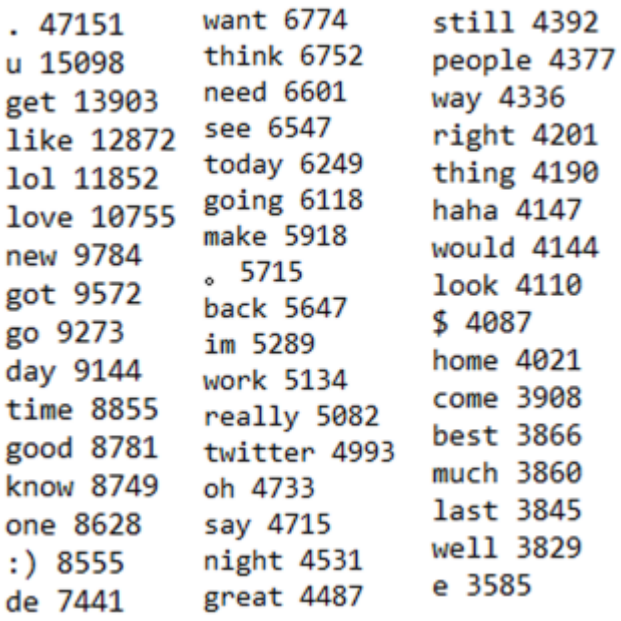


Figure 5.7: 49 Most Frequent General Token Distributions from Control Corpora

As seen in the figure above, some punctuation signs still remain. Even though **“.”** has been removed, what looks like a dot/period, still remains. This is due to the fact that this is another sign or a dot with another unicode, hence it is not being removed.
These general token counts did not lead to any useful insight, so it was decided to apply the positive and negative sentiment dictionaries.

5.6.2 Use of Sentiment

Next, it was explored whether either group used a higher amount of sentiment words in their tweets than the other. To do this, the counted sentiment words (Negative and Positive) from each respective group was compared to the total amount of words in their corpus.

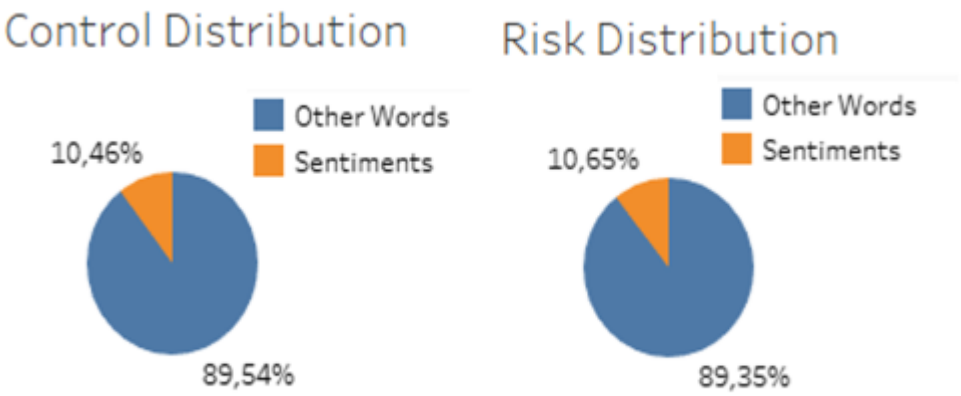


Figure 5.8: Word Type Distribution

As seen, there is a small difference of 0,19% between the two groups, with the Risk Group using a tiny amount more sentiment words than the control group. It is a very small difference, but through the project, the same test was run on different control groups, and every time, the Risk Group corpus would consist of ~ 0.05 - 2,8% more sentiment words, with an average percentage around 1,8%. If this test was to be performed on different risk groups too, the result might differ. Nonetheless, there are insights to be found in these results. This information, did however, not bring insights into what kind of sentiments the two groups are using.

5.6.3 Sentiment Type Distribution

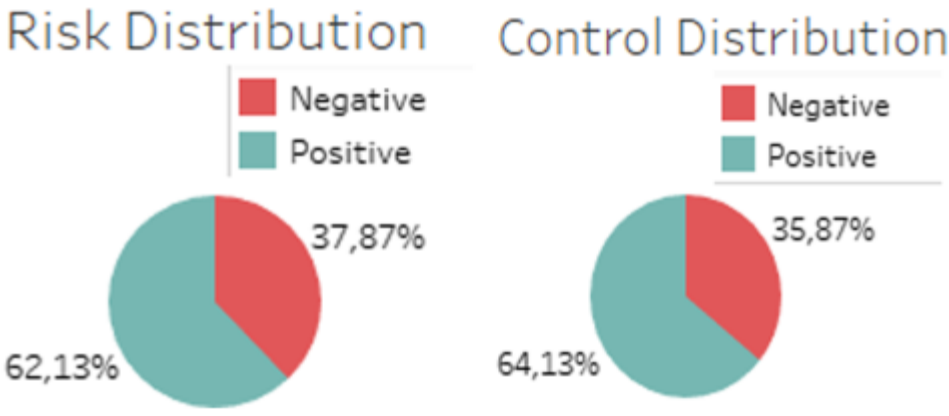


Figure 5.9: Sentiment Type Distribution

The above results are interesting, as they show that the Risk Group has a higher percentage of Negative sentiments compared to the ControlGroup. As before, this test was also run with different Control groups and the results deviated from ~ 0,4% - 5% more Negative sentiment use in the Risk Group. The final randomized control group lies somewhere in the middle with a 2% difference. With the results of the sentiment types, one last step was to be taken before knowledge could be extracted from this information.

5.6.4 Sentiment Word Distribution

As a final product of the preprocessing the Nltk Fdist function was employed once more, but this time, only counting the relative distribution of words found in the sentiment dictionaries. This produced a Negative and Positive count list for each group, making it easier to compare them against each other. By formatting it it this way, R could be used to create lists that would aid with these comparisons.

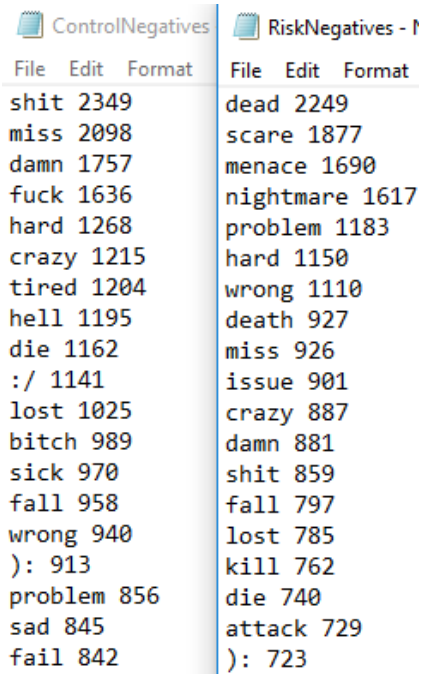


Figure 5.10: Word Distribution List

Some results like the sad emoji "):" is seen in both corpora, while other results differ a lot. Through the control groups used, it was seen that the list did not differ much, and it was mainly the same words to be found in top 30, whereas the words below would fluctuate more. When it came to the Risk Group, words and emoticons like "wrong" and "):" would stick from the initial non-manually filtered groups, while the rest would deviate more compared to the Control Group.

All above Figures are extracted from the final user groups used. These are also part of the results from which insights will be extracted.

5.7 Insights - From information To knowledge.

The data was now converted into structured information, but information alone does not provide answers.

In the context of DIKW, information is perceived as useful, structured and organized data. It is data given context, but no knowledge comes from looking at information alone, e.g. a dashboard without descriptions. To move up the hierarchy and become useful, the information must be contextualised and synthesized, by giving it meaning, as knowledge is created when information is given meaning [Fig 5.2].

To give meaning to the information, it was put into context with the findings from the systematic review, by asking how and why the information is as it is. To further contextualise the findings, Robert Plutchik's **wheel of emotions** and the **NRC Emotion Lexicon** (Mohammad & Turney, 2010) are adopted.

5.7.1 Wheel of Emotions

A human is capable of feeling 34000 different emotions, but such a massive amount is impossible to navigate. Therefore, Dr.Robert Plutchik's research on emotions is applied; work that proposes that all 34000 different emotions stems from eight primary/basic emotion, which servers as a emotional foundation for all others.
The wheel of emotions is used to simplify complex concepts, and to understand which combinations of emotions that can be linked to the identification of SMA.

By utilizing the wheel as two-dimensional it is possible to dive into the emotion wheel and discover which primary emotions(joy, fear, disgust, etc.) can be combined to create secondary emotions (awe, remorse, aggression, etc.).

This is important, as Plutchik's Sequential Model states that, emotions are activated due to specific stimuli, that sets off certain behavioral patterns. (Krohn, 2007; [A6]). As stated, SMA is a behavioral addiction, thus strongly connected to these patterns.

Figure 5.11: Robert Plutchik's wheel of emotions

5.7.2 NRC Emotion Lexicon

The **NRC Emotion Lexicon** is a list consisting of English words and their associations to the eight basic emotions, in combination with a positive and negative sentiment. The list is created by the experts of the national research council of Canada and tagged manually by crowdsourcing on Amazon's Mechanical Turk. The NRC Word-Emotion Association Lexicon is accessed through the Non-Commercial license. The list consist of 14182 words divided into below categories. Note that a single word can be found in multiple boxes.

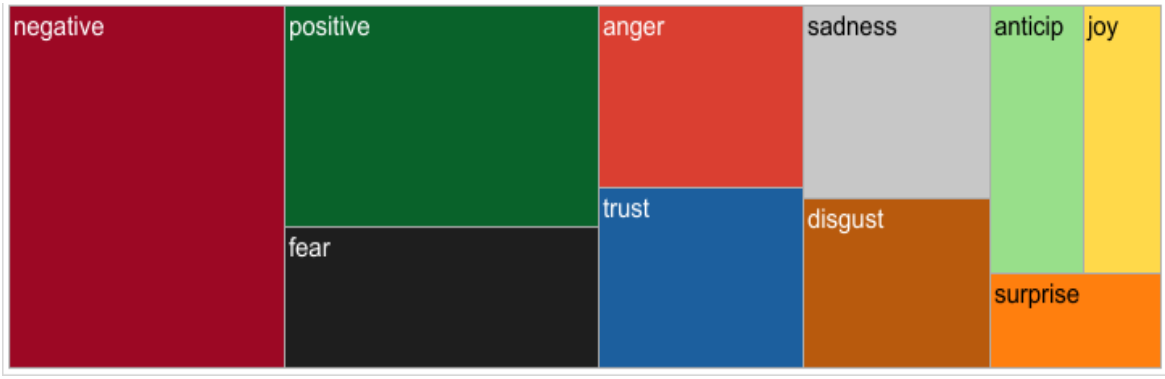


Figure 5.12: Treemap showing number of words associated with each category

5.7.3 Basic Emotion Distribution

Earlier in the paper, there has been a big focus on frequency/time-usage (tolerance), as this was used as the initial proxy. This section however, delves into different factors, that can be used to identify SMA. The 8 basic emotions connected to these factors, stems from the mentioned methods and consists of **anger, anticipation, disgust, fear, joy, sadness, surprise and trust.**

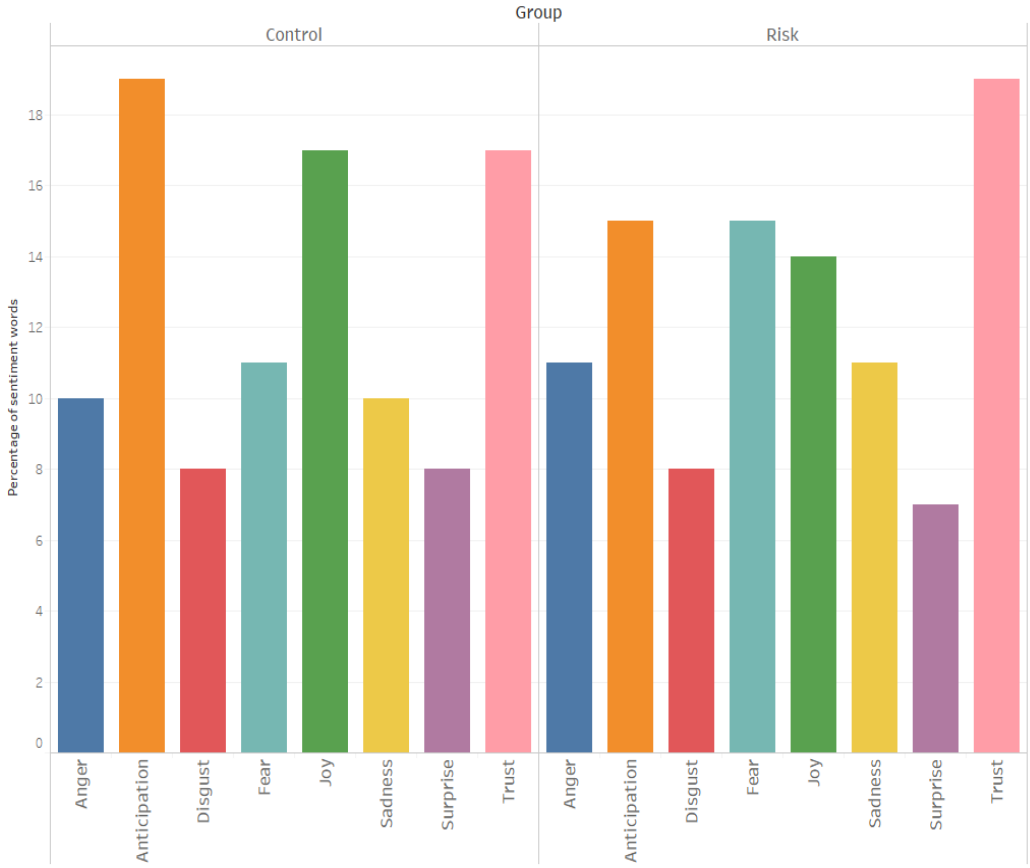


Figure 5.13: Emotion Distribution

Use of sentiment

By making use of the NRC Emotion Lexicon containing 14183 words, the use of sentiments results from section 5.6.2 changes from 10,46% and 10,65% to 18% and 20% respectively. This is almost a doubling, which, taking into account the amount of words in the lists, makes sense. Furthermore, it depicts the average difference of 1,8% better.

Use of Emotions

Figure 5.13 depicts the primary emotions in each text corpora; risk and control group. It proves interesting, as the risk group is far more abounded with negative emotions including; sadness, anger and fear. There is also a larger amount of trust in the risk group, which could be expected according to the literature, as this emotion is connected to bonding and creating social capital.

Moreover, while all emotions are expected in both control and risk groups, the differences between them is of interest, as this can be used to portray the factors mentioned in the literature review to identify SMA (figure 4.2).

5.7.3 Grouping Emotions by identification method factors

Below, the different factors found in the literature to identify SMA, are connected to the different basic human emotions as portrayed by the NRC Emotion Lexicon and Robert Plutchik's wheel of emotions.

Salience (Importance)

Salience is about importance and is displayed as thoughts about spending time on social media to feel important and connected.

Salience is associated with the following basic emotions: *joy, fear, and trust.*
Joy is represented, as a way to feel important to your peers. By sharing everything on social media, it is possible to create a deceptive outward appearance of a life without issues. This can help create a feeling of importance and connectedness.

Fear and *anxiety* often shows together in a person as these symptoms typically overlap. Fear relates to a known and defined threat, whereas anxiety relates to an unknown or poorly defined threat. As fear and anxiety produce similar responses to some dangers, and experts believe that the differences from fear and anxiety based responses, accounts for reactions to various stressors in our environments. Although the words relate to different threats, they are interrelated. When faced with fear, most people will experience the physical reactions related to anxiety. Fear can cause anxiety, and anxiety can cause fear (Sadock et al, 2015). Therefore, fear can be used to prove salience as being anxious about or fearing not being recognized, fear of not being important and/or fear of missing out (FOMO). This could be improved upon, by using more complex emotions

as fear in this case, can also portray other things like the mention of mythical creatures etc. However, it is believed that this connection is important and better left in, than out.

Trust is considered as a way to express salience, by connecting with people and improving social capital, thus creating a bond of trust with your peers.
Social media addicts tend to represent themselves as important to become connected with new peers; thus creating social capital.

The literature expresses that addicts tends to use social media for salience purpose. This connects to results in figure 5.14 with the risk group showing more emotions connected to salience. This creates a clear connection to the literature.

Mood-modification
The utilization of social media to forget personal problems. E.g. to reduce feelings of sadness, anger, anxiety and more.

Joy and trust are used to reduce negative emotions such as sadness and anger, by utilizing positive emotions to do so. The *surprise* emotion, focuses on distractions and new situations to forget personal issues .

The results presented in figure 5.14 shows that the control group use more sentiments connected to mood-modification, as this is connected to positive sentiments. This makes sense as the control group has a bigger percentage of positive words. However, the use of positive words in the control group and risk group are very different. The control group mostly makes use of lighthearted positive words and abbreviations such as **(Haha, Like, <3, Fun, Wow, Lucky, Happy and Cool)** while the risk group has a much larger use of positive words such as
(Hug, Wisdom, Reform, Succes, Interesting, Support, Freedom, Smile, Successful, Courage and kindness)

According to the national health service in England, mindfulness is a useful tool in reducing stress, anxiety and depression[14]. The words more regularly used by the risk group can be associated to mindfulness and self-compassion (Neff, 2012). These words represents ways to deal with personal problems and reduce feelings of anxiety and anger.

[14] https://www.nhs.uk/conditions/stress-anxiety-depression/mindfulness/

It is important to note that these individuals, according to literature, has a higher chance of being anxious, stressed etc. This means that these words do not depict an actual state of mindfulness or self compassion, but instead it serves as a way for these individuals to create a deceptive outward appearance of a life without issues; hence feeling better about themselves (salience).

Conflict

SMA can create real-life conflict, e.g. prioritizing social media over hobbies, family and friends. This has a negative impact, as the addict spends more and more time on social media instead of partaking in offline social activities.

Conflict involves four of the basic emotions: sadness, anger, disgust and fear. When displaying sadness on social media, this is oftentimes a cry for help from the addict. With this display of sadness, the addict hopes that a loved one will reach out. While non-addicts are more prone to reaching out to loved ones in an offline setting, the addicts is prone to do so on social media.

Furthermore, *anger* presents conflict, as it shows resistance to problems and a strong negative emotion about these issues. Anger is shown in demeaning, retaliatory and aggressive behaviors, leading to responses that can start, escalate and prolong conflicts. Furthermore, dealing with anger is often taught as a part of conflict management courses[15]

Disgust, like anger, is a way of expressing aggression. When the target of a moral violation changes from oneself to another individual, anger decreases, but disgust increases. Anger is associated with direct aggression and disgust is associated with indirect aggression (Molho et al, 2017). Both kinds of agressions are connected to conflictual behaviour.

Fear is a cause of conflict. Humans fear the things that are different from us because we don't understand them(The concept of the "other"). This can be said of cultures, religions, races or any group of people. We feels safe inside our comfort zone and we fear the other will change this. Fear is seen as the cause of many communal conflicts

[15] https://www.conflictdynamics.org/anger-and-conflict/

(Lischer, 1999) such as hate-crimes, nazism etc. Furthermore, as mentioned, fear is closely related to anxiety, which can cause conflicts via. fight or flight situations.

The results in 5.14 are expected, as the risk group has a higher percentage of negative words, which leads to the risk group showing more emotions connected to conflict, as negative emotions and words are connected to this.

While the basic emotions can be difficult to attach to these identification factors, this is done to extend on the positive and negative words; thus connecting the results of the data to the literature review to gain knowledge. This knowledge can then be utilized for future work in the field and serve as factors in a classification model, rather than using frequency data solely as a proxy for identifying SMA.

Tolerance

Tolerance is expressed through the time spent on social media, when it gradually increases compared to what was initially intended. This shows the users urge to use social media. Tolerance is represented by the frequency data; hence we do not use emotional words to portray this.

Withdrawal (restless if not utilizing social media) and **relapse** (failing to quit or lower usage), has proved difficult to identify via the data, as these revolve subjects outside of social media. Withdrawal is impossible to prove, as restlessness when not using social media calls for offline monitoring.
Relapse could possibly be identified by monitoring an individual's social media usage in combination with rehab treatment.

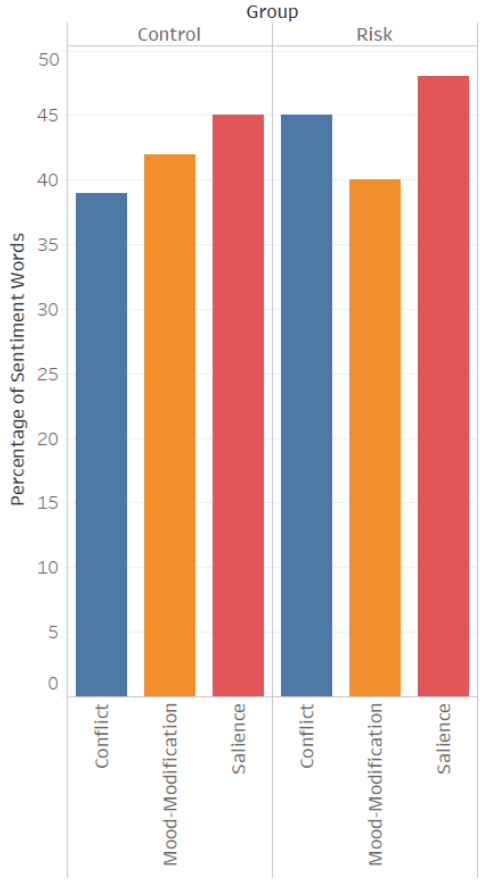


Figure 5.14: Grouped Emotions Distribution

These insights prove that a connection between the literature and data exists. This connection, does not create a tool for identifying Risk groups, but paves the way for future research in the field. It is suggested that these findings are to be used as factors in a future classification model. Furthermore, it is advised that this future model is built on a per-user basis with a hand-labelled data set of interviewed users. Building the model this way, allows for the identification methods from the literature to be a part of the classification model.

5.8 Limitations and Future Research

This study comes with two major limitation categories that can be improved upon for the sake of future research.

5.8.1 Data Collection

For the data, pre-scraped data was chosen. This created the issue that there was not any way of tagging addicts and non-addicts, leading to the use of frequency as a proxy for addiction. Furthermore, this does not assure a complete image of the users, as the tweets might be collected randomly, in a way that does not make sure that all tweets of a given user from the time period exists in the data; something that could twist the results unknowingly.

It is our suggestion, that for a similar study, the data should collected for the specific purpose, and created in combination with the existing methods. It is therefore proposed to start off with a smaller data sample of screened individuals, that would be tagged as control group and addicts from the start. This would allow to skip the proxy and identify actual addicts and not a risk group. The data collection method can thus be improved upon in the same way, and if the study had to be repeated, this is how the data would be collected..

by using supervised classification, the variables could be regressed the found variables against

5.8.2 Data Processing and analysis

The first limitation was a self made one, the text was mine on a group basis, whereas it would be possible to gather more in-depth results, with a per-user corpora. Furthermore, such a corpora would allow for the application of both unsupervised and supervised machine learning. Even with limited knowledge existing algorithms could be applied to the issue if the data was mined per user, as these call for a specification of N-groups in the various classification methods. By running an unsupervised classification, it would be possible to check if there are insights to be gathered in the way it groups the users. As they would be tagged already, the tag would be left out for the purpose of checking the results against the already given tags. Furthermore, by using supervised classification, the

found variables could be regressed against the Addict/Control groups to test different hypotheses, with the goal of testing them with a measurement of statistical significance.
A natural extension for future research would thus be to use the insights from this study to conduct one based on machine learning techniques on individuals.
This method would have to extend other factors such as frequency over time(Testing further for tolerance and relapse). It should also analyse the tweets on a per tweet basis, to give each tweet a sentiment score instead of the collection of words, as this would be more useful in providing context, as a positive word can be used in a negative sentiment sentence and vice versa.

By doing this, it would be feasible to create a prediction model with about 69% accuracy by using machine learning. The statement of feasibility is given with the fact in mind that, in the case of choosing the sentiment of a text, studies show that humans agree on between ~ 70%-79%, making it impossible to reach a perfect computational model (Kruuse et al, 2017). Furthermore, multi-class machine classification of suicide-related communication on Twitter has reached a F-measure of 0,69 on the suicidal ideation class (Burnap et al, 2017). By improving the study with those key points in mind, a big data identification method could be created through the use of predictive analytics.

6.0 Big data Value

6.1 Part introduction

While the previous sections has focused on proving feasibility of the new method, the value of such a method does not lie within what is possible, but instead what is done. This creates the foundation for exploring the possibilities of this new method for identifying social media addiction. While a complete tool has not been developed during this thesis, the emphasis on big data and SMA, has set the tone for future work. Section 5 proves that it is possible to create a classifier to identify SMA risk groups and individuals via twitter data.

Big data has become a way of producing actionable insights and knowledge within businesses. Managers can use big data as a tool to measure key performance indicators and obtain radically more knowledge about their businesses. By directly translating this

into knowledge, it can be used to improve decision making and performance, leading to the emergence of new ways of managing (McAfee & Brynjolfsson, 2012).

"*You can't manage what you don't measure*" (McAfee & Brynjolfsson, 2012)

Big data has reached every sector in the global economy (Manyika et al, 2011). While big data could seem as another word for analytics, four key differences distinguish this: *Volume, Velocity, Variety* and *Veracity.*

The benefits are many, as new knowledge and insights are produced. These can be utilized to automate decisions and processes, which ultimately provides an increase in efficiency compared to manual and subjective "gut feeling" analysis (Manyika et al 2011). This creates a competitive advantage as companies that characterise themselves as data-driven perform better on objective measures of financial and operational results (McAfee & Brynjolfsson, 2012).

Big data has many benefits and potentials, and how to make use of big data is only limited to the imagination of its potential (Beer, 2016). However, the use must be examined critically, as some argue that big data is an invasion of privacy. This raises abounding ethical questions and comes with risks that must be taken into account (Boyd & Crawford, 2012; Davis & Patterson, 2012). Particularly the intensification of digital surveillance, could create consequences for society as many populations are unknowingly extracted from data for different uses (Boyd & Crawford, 2012; Brayne, 2017; Myers, 2017). While these can prove useful for both companies and individuals, such as in the case of SMA, these should be examined thoroughly, as good intentions do not condone wrong use of data. This places serious questions to consider regarding ethical use and corporate social responsibilities.

The remainder of this thesis will focus on discussing the value the suggested identification method can create, while also discussing the ethical considerations which must be taken into account, along with the corporate social responsibilities.

6.2 DIKW from knowledge to wisdom

As mentioned, the value does not lie within the knowledge, but in how to use it. To extract this value, the knowledge must be turned into wisdom and possible decisions. This is done by reflecting on the knowledge (actionable information) found in the results section and how this suggested big data method can create value, turning the knowledge into wisdom (ability to act in any given situation) from which decisions can be made (Rowley, J. 2007).

It is important to acknowledge that the notation of wisdom as knowledge given insight is used, and that while knowledge is based on the past, wisdom leads us into the future, by revealing principles and possible directions. Furthermore, decisions are made as wisdom is given purpose (Figure 5.2).

6.3 Methodology

6.3.1 Data value chain analysis

A data value chain analysis presents a conceptualization of big data analysis, providing distinguishable steps and sorting processes. The concept of the big data value chain will be used to analyze how big data is sourced, refined and given value (Flyverbom & Madsen, 2015).

Flyverbom & Madsen, conceptualizes big data as "*sorting*". They do so as the phenomena of societies and organizations are understood as assembled and fragile. They use *sorting* as an umbrella term for different analytical processes such as classification, categorization, quantification, calculation, valuation and commensuration.

This concept establishes a clear distinction between *data*, *information* and *knowledge*. *Data* is messy unprocessed material such as text. *Information* organized data; grouping data according to some kind of logic like frequency, date, etc. *Knowledge* is information organised in such a way that it creates new insights and value, having an effect, as knowing something means you can act upon it *(wisdom & decision*). This defines data as a type of resource that should be gathered and refined into information, which furthermore needs to be developed into actionable knowledge that can be applied to a specific context.

Based on this, they have created a framework which highlights four distinct analytical moments, regarding the process of turning data into valuable knowledge (Flyverbom & Madsen, 2015).

Production

There are many ways of producing data, but ultimately this concerns human conducts, movement of objects and how these are rendered into quantitative and binary data that can be stored and processed.

Structuring

Structuring concerns the different choices that has to be made regarding databases, classification systems, metadata and how these are ordered and arranged for systematic analysis of the specific context.

Distribution

This analytical step focuses on the access to databases and distribution of digital traces for data-owners and end-users, and how it is negotiated between them. It is of importance to consider this, as many ethical issues could occur if not properly examined.

Visualization

The process of turning data into actionable insights according to the specific context. In this step the role that algorithms play to convert messy unprocessed material (data) into a more interpretable form should be considered (Flyverbom & Madsen. 2015). Here an algorithm is defined as generalized procedures for turning disorganized data inputs, into workable outputs by applying a series of logical rules (Madsen et al, 2016).

In order to gain a deeper understanding of big data projects specific context and its value creations, these analytical steps can provide a focus on the production of knowledge and how data is *sorted,* making sense of the big data phenomenon.

There are two main benefits from using this typology of datafication processes: First, it elaborates practices and interactions involved, creating a more nuanced discussion of big data, with a focus on knowledge production. Secondly it emphasizes how important it is to consider the different processes of *sorting* and technologies used.

These four analytical moments will therefore be used in the specific case of SMA, to discuss the possibilities of actionable knowledge and value creation.

6.3.2 Commercial processes

Data capitalism

A System in which the commoditization of data enables a redistribution of power. Historically, communication and information has been a key source of power, whereas data capitalism results in an uneven distribution that is asymmetrical and balanced towards the actors who have access, and capabilities to make sense of the data. "*This uneven distribution is enacted through capitalism and justified by the association of networked technologies with the political and social benefits of online community, drawing upon narratives that generally fall within the rubric of technological utopianism"* (Myers, 2017).

Commoditization

New ways of collecting and utilizing data are constantly developed, creating value in data, as it can be used e.g. to gather insights and optimize customer purchase behaviour (Myers, 2017). This value constitutes a market that treats data as a commodity to be sold and circulated.

Commercial surveillance

Refers to the way user data is constantly collected for the purpose of commoditization.. While it may not be portrayed directly as a commoditization (selling data), it can be used to improve services such as recommendation systems. The user information and data collected from e.g. cookie technologies are used to improve different types of services, including an aspect of personalisation (Myers, 2017).

The above sets the tone for **content-targeted advertising** and advertisement as a business model of the internet. By utilizing data commoditization and commercial surveillance it is possible to optimize purchase behaviour in an ever evolving online consumer growth, with Google and Facebook as strong examples (Myers, 2017).

These conceptualizations will be applied to the SMA case, to discuss the possible commercial processes from the new suggested method.

6.3.3 Value realization of big data

Günther et al, suggests a value realization method that focuses on both economical and social aspects from big data practices. It does so by emphasizing on three different levels: *work-practice, organizational and supra-organisational* (Günther et al 2017).

Work-practice

A notion of how actors within organizations work with big data throughout their day-to-day interactions. This could e.g. be how to collect and analyze data; leading to discussions about decisions and interactions based on data-driven insights. It is debated whether an inductive algorithmic approach or deductive human-based intelligence approach to big data analytics should be used.

Organizational

This term refers to the structures, norms, resources and procedures the organization uses to coordinate their activities and achieve certain goals. Furthermore, how organizations need to leverage and change their structures, adapt processes and design new business models to realize value from big data.

It is debated how centralized an organization should be to create a hybrid capability structure. Furthermore the business model should be aligned, to make room for change, either as innovation and radical change or as ongoing improvement.

Supra-organisational

Refers to the relations with institutional and technological ecosystems which consists of rival organizations, data providers, regulatory bodies, research institutes, users and customers with whom the organisation interacts. For example an organisation exchanging data with other organizations or customers, for the purpose of mutual benefits.

The above focuses on bridging the trade-off between value and societal implications, transparency and control (Günther et al 2017).

The below figure summarizes the three levels, describing how these are interconnected to create economic and social value.

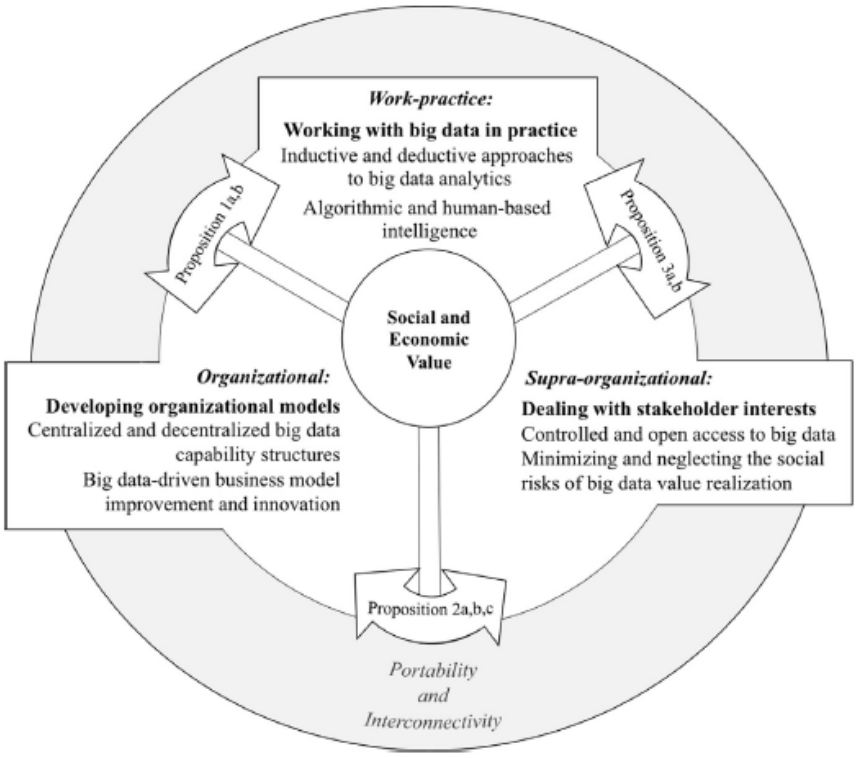


Figure 6.1: Big Data Value Realization (Günther et al 2017)

6.3.4 Business model

A business model is an abstraction of the complexity of a company created by reducing it to the core elements and interrelations. This theses makes use of Uckelmann et al. (2011) business canvas with roots in the work of Osterwalder and Pigneur (2009) due to its proven applicability. The canvas depicts four important perspectives in reaching a value proposition.

The value proposition

This specifies what is actually delivered to the customer. It is defined as the building block, describing the services that create value for a specific customer segment. The proposition serves as the reason to as why customers should choose you over your competition, by solving customer needs or issues (Uckelmann et al. 2011; Osterwalder & Pigneur, 2009). In big data the data, the information and knowledge created are considered as core components of the value proposition.

The customer perspective

Without profitable customers, an organisation cannot survive for long. This perspective focuses on the groups of different people that are reached and served. It includes the *customer segments* that are addressed by the organisation, the *channels* and how *customer relationships* are considered.

Grouping of customers occur, as they are categorized on different attributes such as common needs and behaviour. Examples could be mass vs niche market, segmented vs diversified and multi-sided platforms/market.

The channels refer to how an organisation communicates with and reaches its customer segments to deliver the value proposition. It is basically the customers "touch point" where they can receive news, help, support and make purchases.

Lastly customer relationships is about clarifying the type of relationship an organisation wants with their customers. It is often affected by its channels, determining if it should be loose automated service, or highly engaged service, such as personal assistance (Uckelmann et al. 2011; Osterwalder & Pigneur. 2009).

The financial perspective

The final perspective is comprised of the *cost* and *revenue* structures. The revenue structure represents the sources and ways of revenue generation. Two types of revenue streams can be distinguished: transaction revenues (one-time customers) and recurring revenues (ongoing payments). Ways of generating these streams consists of asset sale, usage fee, subscription fee, advertising and more.

The cost structure describes the most important costs (variable and fixed) incurred while operating. There are different operating methods, each with their own emphasis. For instance cost or value driven approaches which can utilize economies of scale or scope to create a successful business model (Uckelmann, D. et al. 2011; Osterwalder, A, & Y Pigneur. 2009).

Infrastructure

The infrastructure consists of *key partners, key activities and key resources.*

The key partners refers to the network of suppliers and partners that make the business that helps reduce risk and acquire resources. The *key activities* describe the most important actions an organisation must take to operate successfully. The *key resources* are the assets required to make the business model work. These can be physical, intellectual, human and financial (Uckelmann, D. et al. 2011; Osterwalder, A, & Y Pigneur. 2009).

The below figure provides an overview of the business model and how the framework and its different perspectives are connected.

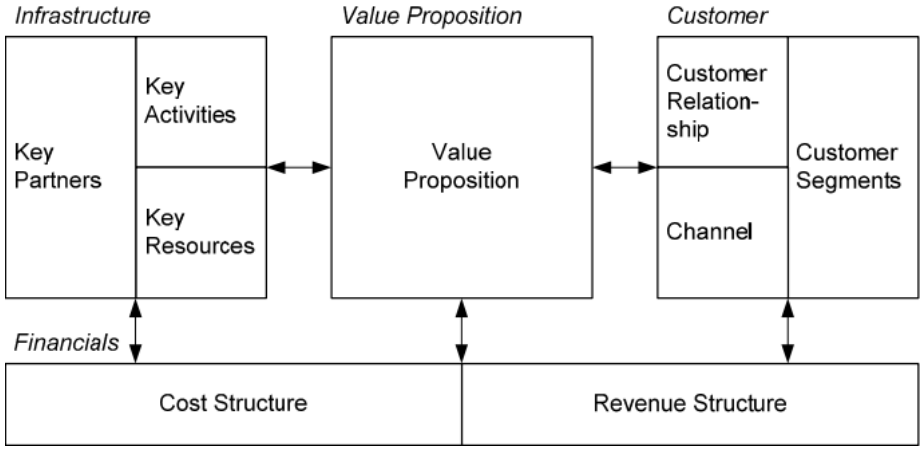


Figure 6.2: Business Model Canvas (Uckelmann, D. et al. 2011)

This business canvas will be used as it has proved favorable in an internet of things (IOT) setting with similar qualities to big data. It is adapted from the work of Uckelmann et al. (2011) as it emphasises the use of data, information and knowledge.

6.4 Data-Driven Business

According to McAfee & Brynjolfsson three of the four V's (volume, velocity and variety) is what makes big data relevant to businesses. The volume of data in today's digital world has reached an astounding amount, as more than 2.5 exabytes (2.5 billion gigabytes) of data is created every day. This fosters endless opportunities for a data driven mindset as data regarding customers, suppliers and operations is being collected non-stop

The velocity (speed) of data creation is sometimes even more important than the volume, as real-time or nearly real-time information makes it possible for organisations to become more agile than competitors, providing rapid insights, such as sales predictions.

The variety plays a big role, as different kinds of data are being created through messages, updates, images, sensors, GPS signals and more. Many of these big data sources are relatively new and as more business activity is digitized, new sources of information constantly presents themselves. It is mostly unstructured data, simply waiting to be released (McAfee & Brynjolfsson 2012).

It is also important to mention the fourth V(Veracity), as false data can be a liability instead of an asset.

Big data is now part of every sector and function of the global economy. The ability to collect and analyse data, has become more accessible than ever before, due to the notion of Moore's Law[16] and its digital storage and cloud computing equivalents. Big data can be used to create value across sectors of the global economy as it brings a tremendous wave of innovation, productivity, growth and new modes of competition and value capture (Manyika et al 2011).

Examples include health care, retail and more. In healthcare, value can be created in many ways, such as automated pricing, predicting diseases (from high risk patients), remote patient monitoring, clinical decision support systems and more. In retail it can be used for inventory management, marketing and to optimize assortment, pricing, placement and design. (Manyika et al 2011).

According to Manyika et al, value from big data can be generated in the following ways:

Creating transparency

Making big data more easily accessible to relevant stakeholders can create tremendous value e.g. by lowering search and processing time, so that the employees that need it can be given access.

[16] First described by the Intel cofounder Gordon Moore, it essentially states that the amount of computing power which can be purchased for the same amount of money doubles every two years

Enabling experimentation to discover needs, expose variability and improve performance

Organisations can collect accurate and detailed performance data and act upon this. The collected data is analyzed to understand root causes, and improve performance, in cases such as product inventory and sick days.

Segmenting populations to customize actions

Allows tailoring of products and services to satisfy needs or solve issues that a highly specific segment calls for. An approach well known in marketing and risk management.

Replacing/supporting human decision making with automated algorithms

Improving decision making by analyzing entire datasets from customers, employees or sensors embedded in products. This can lead to the minimization of risks and discovering of valuable insights which might not have been found otherwise. E.g. automatic flagging of tax evasion candidates.

Innovating new business models, products and services

Innovations the creation of new products and services in an organization, by enhancing existing ones and inventing new business models (Manyika et al 2011).

"Data-driven decisions are better decisions-it's as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management" (McAfee, A. & Brynjolfsson, E. 2012)

Big data does not remove the need for vision or human insight, as **leadership** setting clear goals and asking the right questions is needed. To be a successful leader, being able to spot opportunities, persuade employees to work hard, understand market developments, think creatively and more is needed. Furthermore, the leadership needs to accept that decision making is changing with big data. **Human capital** is required. As data becomes cheaper, making sense of the data become more valuable, as skills from data scientists such as cleaning and organizing data sets, key techniques and visualization can help find the gap between correlation and causation. **Technology** is also a challenge that requires a serious investment in tools to handle the volume, velocity and variety. An effective organisation needs to put information and relevant decision rights

in the same location to optimize **decision making**. Cross functional cooperation should be a focal point, as the people that understands the context needs to work together with the data scientist. This puts an immense pressure on the **company culture** as a data driven approach has to be instilled, moving away from acting on intuition (McAfee, A. & Brynjolfsson, E. 2012).

Despite the challenges, big data can create a competitive advantage as organisations that characterise themselves as data-driven perform better on objective measures of financial and operational results (McAfee & Brynjolfsson, 2012).

6.5 Big Data Value Realization

The target group of the suggested method is rehab centers, as they can benefit immensely from this new way of identifying SMA.

This section is written, as if the feasible tool already exists, as to create a better picture of the value creation possibilities, and use of this newfound knowledge to make future decisions. Furthermore, this section takes a theoretical standpoint as of how to realize value, and does not directly treat this as a case study, but instead with a broad focus on rehab centers.

Note that this section forward, uses RU (Risk users) as an abbreviation of users at risk of becoming social media addicts.

6.5.1 Data value chain analysis

The process of turning real-time data into valuable knowledge consists of four important steps: *production, structuring, distribution* and *visualization* (Flyverbom & Madsen, 2015). These must be considered to gain a deeper understanding of a big data projects specific context and value creation.

Production

The data is initially produced when users interact with twitter in different ways; most importantly, when they write and publish a tweet. The identification method suggested by this thesis focuses on two specific indicators in the produced data. The Frequency of which an user interacts with the social media platform, and the sentiments found within the text represented by the tweets.

The data produced by the twitter platform is a simultaneous expression of the user, and the restrictions put forth by twitter. Such a restriction could be the character limitation of 280 characters (140 at the time the tweets were written), which has proven to have a big impact on the use of abbreviations to prevent hitting the limit. This proves the point of Flyverbom and Madsen, that you are constrained to work under the designs of the interfaces that shape the datafied articulations as *"Data is neither 'found' nor 'raw', but rather produced in different ways in different contexts"* (Flyverbom & Madsen, 2015).

This indicates that data is always manipulated by the context in which it is produced, in this case by Twitter. This presents a significant issue to be considered when utilizing big data and an external data source. If e.g. Twitter wanted to change their API due to commercial purposes, it could change the *production* of data considerably. This could result in technical issues by e.g. disrupting real-time data streams, limiting reliability.

Another issue would be access to the Twitters full stream of data, as it is dependent on the free API or negotiation (price vs time spent and filtering) and hence not guaranteed, as conditions for distributions are set by Twitter. If you do not want to invest in Twitter data, the possibilities are limited[17] (historical vs real-time & amount of tweets etc.). These limitations could prevent the realiszation and optimization of the identification method, by only collecting **some** rather than **all**. Gaining access to data can be difficult and expensive (section 5.1.2) for a *data poor* "outsider", due to the amount of leverage that *data rich* organisations like Twitter have (Boyd & Crawford, 2012).

By choosing a specific type of data for a specific context, you are limited to the restrictions put forth by said data. In this case, choosing Twitter data, influences the sorting process with our bias, as we value this kind of data higher. Ultimately, other kinds of data, could prove better in identifying addicts, but due to our worldviews we have limited ourselves. After the data is sorted, *knowledge* is created and can be used for future transformation into *wisdom* and *decisions*.

This restricts the production of the data for the target group (rehab centers), due to the external sources, the production can be considered partially influenced and dependent on the strategic goals set by the people sorting the data. Only data with an *effect* is

[17] https://developer.twitter.com/en/pricing.html

chosen, providing partial control over the production of data, as the data is chosen to understand aspects in the specific context, by converting unstructured data into structured (Flyverbom & Madsen, 2015).

Structuring

Twitter initially decide the format and available meta-data. Examples could include tweets, retweets, geo-locations, time-stamp and usernames. This represents a technical issue when using the suggested method to identify RU, as Twitter has the sole role of deciding data formats and available meta-data.

During the process of structuring data, it is important to be aware that the chosen sorting strategy, impacts which data becomes visible and valuable in a project, rather than just being a technical issue (Flyverbom & Madsen, 2015).

The people using the suggested method will therefore be the deciding factor, as they make sense of the unstructured data from Twitter, and decide on important metadata to use. In this case authors, and content of the tweets are vital, as this method relies on these to identify RU. Hence, the challenge lies in making Twitter's structure valuable, and not establishing a standard for structuring (Flyverbom & Madsen, 2015).

It would be expected that the people using the suggested method, would only include relevant meta-data, to use this for the identification of RU as well as approaching these individuals. It is evident that choices regarding structure has a significant impact on results and insights. It is therefore an important issue to consider when utilizing big data (Flyverbom & Madsen, 2015)

Distribution

The rehab centers can profit from identification of RU, as they can locate segments more likely to need their services. The rehab centers can then benefit from monitoring their clients data, pre, peri and post addiction, utilizing the factors mentioned in section 5.

There are however, a few challenges. One being the distribution of data from Twitter, as the available data is limited. Furthermore, an issue in the distribution process is that Twitter data continuously changes, making it hard to identify standard guidelines (Flyverbom & Madsen, 2015). Employees at rehab centers might have limited experience

with big data which creates a challenge in managing the data, ultimately calling for investment in human capital to do so. This could potentially result in some rehab centers disregarding the project, as they do not want to become data-driven or have the capital to do so.

Visualization

Determining RU requires an algorithm for successful identification. Such algorithm can be developed by using the evidence found throughout the thesis in combination with the suggested factors in section 5. These factors allows for identification and extraction of the subjective meaning and emotions from text, making it strikingly relevant to identify, monitor and analyze social media texts (Lassen et al,. 2014).

The algorithms can therefore be seen as more than technical achievements, given the human and organizational choices that lie behind them. Subjective choices decide which analysis to perform in the specific context and are therefore based on subjective interpretations, such as the results portrayed in section 5.

Grounded on these automatic algorithmic operations, visualisations can be made which give insights into RU. These can be used to both identify and monitor the individuals, either real-time or periodically. It is of importance to note how the continuous monitoring of the clients could prove as an attractive service, both for the individuals and analytical purposes of the employees in rehab centers. Specific implementations of these visualizations however are to be discussed in future work.

Flyverbom & Madsen suggests that big data neither proves neutral nor objective observations, but rather context specific insights which can be acted upon. This therefore proves how vital it is to understand the data value chain and the process of sorting unprocessed data into valuable knowledge, to understand the value that a big data project can provide for decision making.

6.5.2 Commercial processes

Once the data is obtained from Twitter and sorted into valuable knowledge, it must be considered how this can be used to improve commercial processes for the target group. A rethinking of rehab centers business models towards a more data-driven approach is needed to reap the benefits of the new identification method. While this would not

change the rehab centers completely (their treatment and services), it would be an incremental change in adopting a more data-driven approach, to identify and connect with individuals in need of your service.

Identifying SMA risk groups

The commoditization of data at first occurs from Twitter, as rehab centers would need to invest in the required twitter data, to optimally use the method suggested in this thesis. However, being able to identify a certain segment of people, whom are at risk of becoming social media addicts, could provide immense benefits..

The commoditization of the data produced by rehab centers using the suggested method can therefore not be reflected upon as regular commoditization (selling the data) but rather, using it to optimize buying behaviour of possible clients and internal processes. The commoditization would therefore happen indirectly, as rehab centers would be able to identify a new customer segment within Twitter. This could help boost the number of clients that turns to treatment at rehab centers, instead of going untreated. This is done, by quickly reaching the affected individuals and motivate these to receive treatment; thus improving the rate of treated addicts. The market for addiction treatment is estimated to be $35 billion a year[18], and an innovation of the business model within such a big market, has the potential to create economic value.

An important issue however, is that this method is only representative of Twitter; hence only applying to the users of this specific site. Furthermore, because of the markets size, for-profit focus[19] and untapped market of addicts that do not use treatment centers (just under 11% of addicts gets treatment), there is an opportunity to invest in technology and big data and utilize the new method. Depending on the implementation, it is possible to add a data-driven approach to the business model of treatment centers.

[18] https://www.forbes.com/sites/danmunro/2015/04/27/inside-the-35-billion-addiction-treatment-industry/#207fcb2317dc

[19] https://www.rehabs.com/pro-talk-articles/the-demise-of-for-profit-addiction-treatment/

Approaching SMA individuals

The big data method can also be used to approach RU. When they are identified on twitter, rehab centers will be able to approach these faster than ever before.

By using big data to identify RU, two new options for approaching potential clients present themselves: targeted advertising and direct messaging. As the risk groups are identified these are all listed with their usernames. Twitter offers the possibility to reach a specific target group based on various filters (usernames, location and more)[20]. Thus creating an opportunity to reach your potential clients better, by focusing on a customer segment consisting of RU. This does however, create a monetary challenge as advertisement on Twitter is expensive.

Furthermore, advertisement could be done by creating a Twitter organizational user that has a live chat. In this way, advertisement through direct messages to the individual users can happen as soon as they are identified. This also creates a more personal approach to the potential customer, via this new channel.

Using big data pre, peri and post treatment

By being able to use commercial surveillance, the progress of customers can be monitored. Not only can rehab centers monitor potential RU and identify these easier, they can also use the data to monitor progress pre, peri and post treatment. This could ultimately improve treatments and services offered by rehab centers as new data is produced regarding the success rate of treatment, potentially creating wisdom that can be used for future decisions. Furthermore, within the data, time-stamps of the tweets is available and could thus be used to optimize opening hours for the direct messaging live chat, that could be implemented to approach RU individuals at optimal hours. However, future work should investigate this further.

Overall the suggested method can potentially use data commoditization and commercial surveillance to optimize purchase behaviour and internal processes within rehab centers.

[20] https://marketing.twitter.com/na/en/solutions/create-engagement/audience-targeting.htm

Overcoming the privacy barrier

To reach the potential value, ways of overcoming the privacy barrier must be considered, as offering personalized services based on big data raises concerns regarding identity theft, discrimination and more (Myers, 2017; Brayne, 2017; Davis & Patterson, 2012; Boyd & Crawford, 2012).

To achieve this, there are three ways of convincing customers to overcome their concerns about privacy and reap the potential value (Myers, 2017).

The first consists of *giving things away*. It does not refer to actually giving physical items away, but rather a focus on value of the free (Trial approaches) and having an open network, to obtain ubiquity. In a network economy "*Ubiquity drives increasing returns"* (Myers, 2017). The most cost-effective way of creating economic value would therefore be to make the information freely accessible to your customers. This can be done in several ways such as including it in the subscription fee or service fee from the rehab centers, which could lead to higher success with a data driven approach, making customers focus on the improved service rather than invasion of privacy.

Second, to make data capitalism work, the *potential to create productive intimacies between man and machine* are of importance (Myers, 2017). The data should be used to tailor personal experiences to the customers, including them in the ways it is used. Furthermore the data should be used to improve services offered by rehab centers to clients. Due to personalisation, customers might be more incentivized to allow surveillance and usage of their data, despite the privacy concern. For example, by using the data to monitor progress pre, peri and post addiction, and by allowing customers to access to the insights.

Third, data capitalism relies on technocratic value of data and its potential to augment consumer power (Myers, 2017). Transparency therefore becomes an important feat, as this allows consumers more power over their data, while allowing organizations to produce more data. An issue becomes imminent as data assymetri arises, allowing the power to lie with the organizations and people with technical and economical resources to render data actionable (Myers, 2017).

By considering how to include the above "*Users are placed in a double bind, caught between desires for privacy and the ability to form meaningful communities*" (Myers, 2017). Essentially it is an attempt to weigh the social desire of the customer higher than their concerns for privacy.

The suggested method can be used as a competitive advantage, providing valuable information for rehab centers and their customers. There are however a few issues which must be thought of, to maximize value creation.

6.5.3 Value realization in big data

This section will focus on how to realize value in the three aforementioned aspects: *work-practice, organizational* and *supra-organizational*. It will do so by considering the different challenges of realizing the value within the different levels.

Work-practice

It is often difficult to foresee which insights can occur from the various data sources, due to the variety and granularity. Especially in this case as the data source, Twitter, could be seen as a company with an attitude of collecting data, without a pre-defined purpose, thus promoting what would be called a bottom-up inductive approach to data collection. As a consequence, data is not always produced and collected for the same purposes they are eventually used for (Günther et al 2017). You could define this approach as first seeking data and only after collected, seeking theoretical explanations about a certain phenomenon.

The benefits of using this approach lies in the power of automatic big data analytics. It requires data scientists who have the capabilities of various different techniques within machine learning to e.g. use clustering for categorization and finding computer-detected similarities which were not considered before (Günther et al 2017). An issue with this approach in this particular case however, is that using e.g. cluster the data set, there is a possibility that this would not prove useful for the identification of RU, possibly finding unrelated patterns, due to the inductive approach from Twitter regarding data collection.

The deductive approach is much more present in this case, as a clear goal was set for a certain phenomena, testing hypothesis to arrive at a conclusion. Essentially following a an approach which starts from a general theory and then uses particular data to test it

(Günther et al 2017). In this case, to collect, process and eventually visualize RU. While the value generated from this approach is more acknowledged, compared to the inductive, confirmation bias can occur and must therefore be considered.

Balancing the two approaches, could provide benefits (Günther et al 2017). These can be reaped by allowing inductive techniques to some degree of freedom, letting data scientists manage data to perhaps discover unknown patterns and thus innovate further within the field. By setting boundaries and ensuring a degree of deductive approach, realizing business value is more probable in the long run.

Due to the above, algorithmic and human-based intelligence also becomes a matter to discuss. It is important to consider using the algorithmic intelligence (inductive) if possible and balancing this with the human-based intelligence (deductive). The new tool could be considered an algorithmic intelligence that automatically can identify RU. However the importance of human intelligence cannot be stressed enough, as this is required to examine data, discover patterns, derive,- employ and refine insights (Günther et al 2017). Especially concerning this phenomena and its grounding in addiction theory and social science, as the insights found in the Twitter data was made possible due to the systematic review and domain expertise. Data exploration by human operators is therefore of importance, to define a certain phenomenon to investigate, and to test a hypothesis via data. Both of these do have benefits, but it is important to consider meaningful ways of utilizing and balancing these, to allow for sufficient innovation and realize value in specific contexts (Günther et al 2017).

Organizational

Big data can be used in many ways of data-driven decision making, including: accessing, tracking, collecting, managing, governing, processing and analyzing (Günther et al 2017; McAfee & Brynjolfsson 2012). In order to realize value it is of importance to develop, mobilize and use technical and human resources related to big data. To do this, a balance between centralisation and decentralisation must be considered.

Development of analytics *competency centres* to centralize the resources, can be used to deal with shortage of analytics skills and has proven effective (Günther et al 2017).
While such centres can ease adaptation of data governance, it creates an issue in the connection with business units, as it centralizes the data and its insights to one analytical team, without specific expertise within that certain area. The notion that one chief

officer should oversee all data, creates communication costs, need for synergetic collaboration and expensive technologies, skills and knowledge (Günther et al 2017). Especially in this case, that is grounded on its own addiction theories, not related to domain expertise of a data scientist, it could create issues reducing value realization.

While big data in essence is not decentralizable due to security issues, above issues provides reason for doing so. A success factor within big data value realization is the inter-combination of multidisciplinary teams (Günther et al 2017). In this case, the data analysts would need the expertise and knowledge within the theory of addiction, in order to maximize value realization, providing proper assessment of the data, hence establishing the opportunity to analyze and refine insights better. The analysts must therefore truly engage in the business to realize value.

For the reasons above, a hybrid of centralization and decentralization would be recommended to realize value. Both security and expertise within the field should be considered. A way of doing this, could be by centralizing policies in information governance, allowing the analysts to "hire" colleagues for specific situations, to utilize their expertise within the field of addiction.

This can ultimately be used to improve the business model by using a data-driven approach. For incumbent organizations this means re-thinking your current business model and how big data can affect this.

Essentially what needs to be considered is data sources and techniques which can be used to improve existing processes in terms of efficiency and effectiveness (Günther et al 2017). In this case a big data method that can identify RU, leading to new way of identifying a certain customer segment and ways to use targeted advertisement on these, while also offering a new way of monitoring progress of customers. Implementing these incremental enhancements allows for the rehab centers to function in the same manner, while increasing efficiency in attracting customers while also being able to monitor progress of customers and analyze their processes. The extent to which organizations adopt big data and whether they use this for enhancement or innovation therefore relies on the industry and size of the organisation (Günther et al 2017).

Supra-organisational

Apart from organizational boundaries, relevant stakeholders includes data providers, users, customers and more.

To benefit from big data, organizations need effective data exchange between their network of partners while engaging in practices of data disclosure (Günther et al 2017). This opens the debate for whether to have controlled or open big data access. There can be many reasons for why you would want to control your data: privacy, security concerns or when it is considered as a source of competitive advantage.

The initial collected Twitter data is publicly available. However, sharing the data after the sorting process is not desirable for the rehab centers, as this would ultimately share their source of competitiveness, while creating security and privacy issues. This transforms the data into a controlled resource, as algorithms and knowledge lies within the organization.

Open/public big data access is emerging as data is made available to organisations and consumers, as in the case of Twitter. This is in order to stimulate innovation and provide transparency (Günther et al 2017).

The data production is however affected, as it is dependent on the IT infrastructure of organisations and governments, as the data can be deliberately modified. This would lower the reliability and quality of data, e.g social media organizations as Twitter, may not be transparent about how the data is processed beforehand. Hence, a clear agreement between network partners is considered crucial in data collaborations. Consumers and users that are both data sources and recipients of data-based products should be considered as network participants. To use the suggested method, rehab centers would have to deal with legal and ethical consequences of sharing and using data, e.g. having to explicitly ask for user consent, GDPR and Cookie law.

Rehab centers would have to consider how they handle the data, as “new” data is created from the sorting process. The identification and tagging of illnesses in individuals is considered health data, and thus strictly regulated. Some responsibility lies with Twitter, as users can make their Twitter accounts private; should they not wish to have their data made publicly available..

The question then is, how do you minimize and/or neglect some of the social risks of big data value realization. Realizing the value from big data comes with social risks as the data used can be of sensitive concern at risk of being released. Even though surveillance may lead to improved public safety, it also impedes the individual's feelings of freedom, privacy and autonomy (Günther et al 2017). As addiction treatment is within the medical sector, strong regulations concerning big data can be expected and must be considered, specifically regarding the tagging of individuals and monitoring of progress. While some organizations attempt to neglect the social risks to realize big data value, this is not possible due to the massive resistance it could create.

The idea is to find a middle ground where the customers themselves can opt out of being monitored, or decide to receive the social value created by their data (Günther et al 2017). Furthermore, to create value from all three levels, it is of importance to assure alignment between them (Work-Practise, Organisation and Supra-Organisational); this also assures portability and interconnectivity (Günther et al 2017).

6.5.4 Business canvas

The business canvas will be used to create an overview and proposed business model for realizing the value that our proposed method could create.
The proposed business model, contributes with a new perspective on how rehab centers can realize said value by describing the thought process that should go into the incremental change, when using the suggest method. It will therefore not venture into rehab centers current business model, but rather create one that points towards a data-driven rehab center.

In order to create a business canvas (figure: 6.2) four things have to be considered: Value proposition, Customer perspective, infrastructure and the Financial perspective (Uckelmann et al. 2011; Osterwalder & Pigneur. 2009).

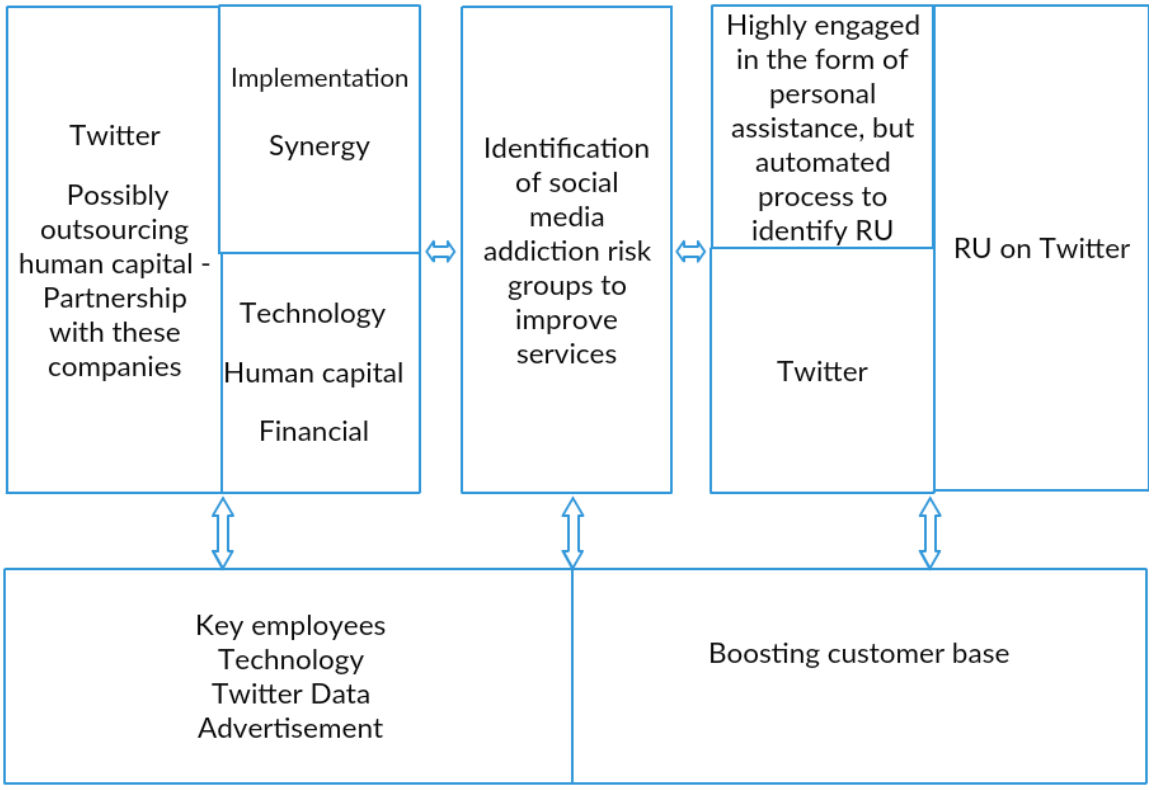


Figure 6.3: Data-Driven Rehab Center Business Model

Value proposition

The suggested method brings an overall value proposition consisting of *improved services through effective identification of RU.* By being able to identify RU, it allows rehab centers to effectively focus on a group of people which might be at risk of addiction and thus in need of their services. For the rehab centers, this could mean an increase in customers, as targeting these becomes much easier with the insights gained from big data. This, in combination with big data surveillance for monitoring of progress of their clients, provides a tool which can help customize their services to each client, while possibly improving treatment. It essentially allows the rehab center to tailor a service directly to the RU and creates a convenience for these as a new service of Twitter direct messaging can be used to gain information regarding the treatment as well as consulting.

The value that using the suggested big data method delivers to the customer, is therefore an overall better service, by being able to identify and reach the RU faster than before.

Customer perspective

As mentioned in the value proposition, being able to identify RU on Twitter, creates an opportunity to identify and reach a specific *customer segment* faster.

It creates a new segmented market, as rehab centers are able to reach possible customers with another channel (targeted marketing) and convince these to utilize their center.

A data-driven approach would use social media (Twitter) as a new channel to reach customers, as well as conducting digital marketing within. This will allow for a "touch point" to communicate with customers who can receive news, offers, support, raise awareness and more.

This sets the tone for the *customer relationship,* as the data-driven model suggests a highly engaged service in the form of personal assistance regarding identification and treatment of addicts. The rehab center should thus focus on customer acquisition through this new channel; resulting in a boost of sales (Treatment). While this approach, creates an automated service in the sense of identifying potential customers, it calls for rehab centers to use these insights to create a personal service and experience.

Infrastructure

The *key partner* in such an approach would be the data organization (Twitter), as this is where the data is produced. Reaching out to these and forming a partnership could help reduce risk and uncertainty towards the data needed from this external data source. Being able to form a partnership with such a big company could however, prove difficult, as they would probably be profit focused. A "Customer" approach to gathering data, thus seems more likely, as this data is commoditized. Furthermore, to avoid a hefty investment in technology and human capital, outsourcing these could be considered. This would create key partners in the outsourced organization.

The *key resources* of data-driven identification of RU comes in three forms: physical, human and financial.

The physical focuses on the technology that would be needed, should a rehab center implement big data. Internal storage or external database management would be needed,

additionally, running the classification in real-time would require a large amount of computing.

The human resources needed would be people that are proficient in data analysis and how to realize value from this. Human capital is therefore needed in order to realize value, as employees within a rehab center most likely does not have experience with data analysis.

Lastly, financial resources are needed to acquire the mentioned human and physical resources.

The *key activities* relies on the above, as the implementation of a data-driven approach is required to realize the value proposition. The implementation of this allows for a problem solving method, as they identify their customers better and can improve their services. It calls for continuous training for the rehab centers regular employees (with expertise in the field of addiction) to work side-by-side with the human capital that has expertise in the data domain.

Financial perspective

Revenue streams can be considered as transaction revenues resulting from one-time customers, assuming it is in the rehab centers best interest to give proper treatment to their clients and their addictions. While there will be some recurring payments from customers relapsing, it will be assumed that transactional payment will be the most common in the form of fixed pricing for certain treatments, while some are volume dependent as they differ according to the period of the stay[21]. A data-driven approach can create two new revenue streams. Firstly, it can increase the fixed price for receiving treatment as the improved service can be priced higher. Secondly, by identifying a new customer segment and being able to target these better, they can increase their customer base, thus increasing revenue from transactional sales. In the specific setting it must be considered if an increase in customers is enough to cover costs and create value or if an increase in price is also needed and/or possible. Other ways of creating revenue with big data could be a usage fee to see progress through visualizations of big data, however, this does not create a feeling of "giving things away" and "transparency" which is considered needed to convince customers to think less about privacy.

[21] https://www.addictioncenter.com/rehab-questions/cost-of-drug-and-alcohol-treatment/

The *cost structure* proposed, is a value-driven cost structure, with less focus on the minimization of costs but rather a premium value proposition with a high degree of personalized service. Costs to implement and maintain this business model consists of key employee(s) for data analytics and customer service via the new channel (human capital), technology to make use of data, the data itself and advertisement costs for the targeted marketing. These costs are necessary to consider when implementing the data-driven approach.

It must also be noted, that each specific case is different, and implementation process in different rehabs centers will therefore vary and should be handled situationally. The business model provides an overview of the considered perspectives in realizing value from the suggested identification method.

6.6 Ethical Considerations

While big data is neutral in its ethical nature, it does not provide any guidelines of what is right and wrong, raising a number of ethical issues. This creates a focus on rights and interests, treatment of personal data and how to be a responsible organization, (Davis & Patterson, 2012). The ethics of big data is therefore a very personal topic, situated in specific contexts, where each person has their own thoughts and limits toward the usage of it. While organisations want to benefit from the use of it, they have to be aware of the ethical considerations, utilizing it in a manner that does not have a negative effect on their customers, e.g. violating their privacy.

These questions regarding data and its usage are often asked without clear answers, as many countries have their own regulations and laws when it comes to information privacy laws. Who owns the data and when the gathering of information turn from helpful to harmful are important questions to consider (Davis & Patterson, 2012). Despite companies informing users on how they use their data, a lack of transparency often occurs. A recent initiative taken to regulate the data market is the European General Data Protection (GDPR). This is a data protection and privacy regulation for all people within the European Union and European economic area. The aim is to bring back control of personal data to the individual, rather than the companies utilizing it at will. It regulates the export of personal data from users within these areas, as it simplifies

the regulatory environment by stabilizing data assymetri. Each organization must therefore consider the regulations affecting them.

By using big data in this context, sensible health data is generated. While it may not be generated directly on Twitter, its possible use on clients in the rehab centers, could be defined as health data, as it can be used to monitor progress of social media addiction. Therefore an ethical issue arises, as introducing the method to identify RU, could require informed consent. This should be considered, as health service providers need to protect data from sharing and misuse, as privacy laws often abide this (Syagnik et al, 2017). This implies that managing this type of data, requires a large focus on cybersecurity the compliance of customers to use the data for these exact purposes, e.g. monitoring addiction progress or other business functions. In this case, this is partly avoided throughout the identification, as public Twitter data is used. However, after processing the data, it transforms into possible health data, applying new regulations.

Furthermore, *"Just because it is accessible does not make it ethical"* (Boyd & Crawford, 2012). When creating health and personal data, it is important to act responsibly and ethically. While research within this data could provide benefits, it does not always make it justifiable, thus reflecting upon the ethical use, is very important. The questions is, what does it mean for somebody to be analyzed without knowing it? And making sure it is not used in such a manner, that it could bother potential clients. This should create a focus on *accountability* and multi-directional relationships with superiors, colleagues, participants and the public, to establish professional guidelines that emphasizes protecting informants rights and well-being (Boyd & Crawford, 2012).

Moreover, according to Brayne, another issue arises as the implications for individuals and society are strengthened through intensification of surveillance. This involves the collection, recording and classification of information about people, processes and institutions. She states three things to consider and justify accordingly:

First, why was big data surveillance adopted? As a rehab center, the institutional *goals* that were intended to achieve should be considered from an ethical perspective.
This is vital as the organizational need for surveillance is one of social sorting "*Surveillance today sorts people into categories, assigning worth or risk, in ways that*

have real effects on their life-chances . . . it is a vital means of sorting populations for discriminatory treatment" (Brayne, 2017)

Second, how is big data surveillance conducted? This regards the *means* used to collect and analyze data. It should be reflected upon whether or not these methods are beyond the institutional environment of which it is utilized in. While this method is concentrated on staying within the environment of rehab centers, it is possible that it could be part of unethical conducts, such as selling the data after the classification. Making sure the *sorting* stays within the institutional environment is therefore of importance.

Lastly, what interventions are made based on big data surveillance, and to what consequence? This questions the *ends* and exactly what the key actors do, based on the insights from big data surveillance. The usage of the data should be carefully thought of and explained, while considering the possible consequences arising from this use and managing these properly to avoid harmful forms of surveillance (Brayne, 2017).

The emergence of *data capitalism* provides another challenge as organisations seek profits and market control by collecting information and predicting human behaviour. This becomes an issue as *"Access to data, and the ability to transform raw data into useful information, is asymmetrical, and the power lies in the institutions with the technical and economic resources to render it intelligible"* (Myers, 2017). Data capitalists has the power to transform data, and render it intelligible in many ways, creating data asymmetry. Surveillance capitalists have extensive privacy rights and accordingly many opportunities for secrets. These secrets are progressively being used to deprive populations of decisions in the concern of what information about their lives remains private.

This emphasizes just how important transparency is, to make sure the users do not feel violated. It creates a struggle for transparency as users inner needs and social desires constructs new digital divides (Boyd & Crawford, 2012). User data is produced in a highly context specific situation, where users would possibly not give permission for other uses. Despite this, multiple ways of gathering and storing their data occurs for future use. Data capitalism therefore becomes an issue for some potential customers, as

they are not aware of the many uses, profits and other gains, acquired from the information they surrender.

The new surveillance practices and transparency introduced from big data has developed fear and limited trust in the digital domain. A concern of lack of protection against privacy-breaching activities have led to skepticism, creating a barrier for consumers (Flyverbom, 2017). Particularly this case, involving sensitive health data. This should be considered, as it establishes certain challenges to attract customers. Being aware of this as an institution, makes it possible to implement ways to make customers disregard the privacy issue, making them more reluctant to acquire a new service.

Furthermore, because such sensitive data is handled, security is of great concern as data breaches are becoming larger and more frequent than before (Flyverbom, 2017). This impacts the trust of the digital domain negatively, and creates larger concern for potential customers. It is important to consider which actions to take, should a breach occur. This is in order to avoid unknown people exploiting health data that could ultimately have negative consequences for involved parties (Flyverbom, 2017).

Accordingly, clear protection and regulation of privacy is of great importance, creating a general need for data control (Mayer-Schönberger & Cukier, 2013). The importance for sustainable and responsible data chains presents itself as this is where big data is turned into valuable insights (Flyverbom, 2017).

Because of the above, various topics should be considered when managing data, as both individuals and organizations have legitimate interests in understanding how big data is being managed. As defined by Davis and Patterson (2012), the focus lies on four main topics: *Identity, privacy, ownership and reputation.*

Questions regarding online and offline identity, who should control access to the data, who owns the data, rights, obligations and trustworthiness of data, should be asked.
By not explicitly and transparently evaluating ethical impacts of the collected data, organizations can risk diminishing relationships with customers. Benefits from thinking ethically about your data can include: faster consumer adoption by reducing fear, increased pace of innovation and collaboration, social perks and more.

According to Carrol, *corporate social responsibilities* (CSR), can be applied in businesses, to act ethically and in the best interest of all stakeholders. The focus in on the interaction between society and businesses, and how both can benefit from these, as businesses should aim to act as socially responsible as possible.

He expresses four key areas in his pyramid, which must be contemplated upon: *conomic, legal, ethical and philanthropic responsibilities*. As the ethical responsibilities goes hand-in-hand with previous discussion, the focal point is the remaining three areas.

The *economic responsibilities* regards the production of goods and services that societal members need and profit from. However, the degree to which you profit as an organisation advertises responsibilities. Maximization of profits is a term often used, and while some organisations can benefit from this, it grants consequences for the consumers.

The for-profit focus on the rehab centers has been criticized as they profit by preying on people in n desperate need of their expensive health service[22].

For the same reason, when implementing the suggested method, it should be thought of, whether current profits allow financial room to exclusively improve services, or if this should be used to generate more profit. To be a responsible organisation it is important to be consistent with expectations of societal codes and ethical norms to create a fair image.
As a organization you must obey the law pursue your *legal responsibilities*. In this case, by ensuring that all applicable regulations are met. This must be thought of when implementing big data to create a "social contract" between business and society where both benefit. The idea is to create fair operations that by providing goods and services that at least meet minimum legal requirements.

Lastly, to be a responsible organization, *philanthropic responsibilities* must also be considered. This involves actions made in response to society expecting, that business are "good corporate citizens".

[22] https://www.rehabs.com/pro-talk-articles/the-demise-of-for-profit-addiction-treatment/

For rehab centers this could e.g. include using the new method to create welfare (free treatment), or donating towards more research into the field. This shows society that you care about the well-being of your customers and society as a whole. This is highly desired (somewhat expected) by society, but it does not categorize organisations as unethical should they not do this. For the same reason, this is more discretionary and voluntary and depends largely on whether or not the organisation can and will, provide it. This is deemed less important than the other three areas, as this is considered *"icing on the cake"* (Carrol, 1991; Carrol, 1999).

By considering these three areas as well as the ethical issues mentioned, it is possible to establish a fair image, creating both social and economic value, while also avoiding harmful consequences for organisations and customers. Ultimately, organizations should consider how to align their values with those, which the data represents.

7. Conclusion

Data Analytical Outcomes

The insights provided by analyzing the twitter data paves the way for future research in the field, and proves that **it is possible to identify users at risk of becoming social media addicts with big data analytical methods by extending on the findings in this thesis.**
The key takeaways from the data shows that the risk group has a higher use of negative sentiment words. Additionally, their tweets include more words from certain basic emotional spectrums such as *fear*, *trust*, *sadness* and *anger*, which can be linked to risk factors such as Mood-Modification, Salience, Tolerance and conflict.

We advise that a future classification model is built on a per-user basis with a hand-labelled data set of interviewed users. Building the model this way, allows for the identification methods from the literature to be a part of the classification model. Furthermore our research suggest, that it would be desirable to factor in withdrawal and relapse signs in addition to timeline data.

By using our research to create a machine learning classification method, we believe it would be possible to create a prediction model with about 69% accuracy by using machine learning.

Value Realization outcomes

Research proves that the suggest method **can create value for business and society, as it is possible to gain a competitive advantage, and improve the quality and quantity of treatments (possibly resulting in less addicted individuals), by using the suggested method.** However, different issues must be taken into consideration.

First, the data value chain analysis has proved that many challenges arise along with big data. Big data projects are defined by their specific context. This makes it crucial to understand data value chains and their processes in order to realize value.

Second, in order to use data capitalism to create value in this specific context, the potential commercial processes of the suggested method must be considered. By identifying social media addicts, new ways of customer segmentation, approaching customers and monitoring progress of these are made possible. However, it is important to note that certain privacy issues must be handled to maximize the value realization of such identification.

To realize value of a big data project, it must be considered at work-practice, organizational and supra-organisational levels; making sure that your organisation is aligned with these. The findings shows that there are many challenges to consider, when guiding future decisions for changing the way we identify and work with social media addiction and data-driven business models.

Working with big data presents many ethical and privacy concerns, emphasizing corporate social responsibilities and how to realize value from big data in a responsible way.

Working with big data presents many ethical, privacy and corporate social responsibility concerns, which must all be dealt with, when pursuing the realization of value from big data in, a responsible way.

8. References

Sarah Kenyon Lischer (1999) Causes of Communal War: Fear and Feasibility, Studies in Conflict & Terrorism, 22:4, 331-355, DOI: 10.1080/105761099265676

Molho, C., Tybur, J. M., Güler, E., Balliet, D., & Hofmann, W. (2017). Disgust and Anger Relate to Different Aggressive Responses to Moral Violations. *Psychological science*, *28*(5), 609–619. doi:10.1177/0956797617692000

Abbasi, I. S. (2019). Social media addiction in romantic relationships: Does users age influence vulnerability to social media infidelity? *Personality and Individual Differences,139*, 277-280. doi:10.1016/j.paid.2018.10.038

Abbasi, I. S. (2019). Social media addiction in romantic relationships: Does users age influence vulnerability to social media infidelity? *Personality and Individual Differences,139*, 277-280. doi:10.1016/j.paid.2018.10.038. Retrieved from: https://www.sciencedirect.com/science/article/pii/S0191886918305798?via%3Dihub

Ackoff, L. R. (1989). From data to Wisdom.

Akdere, M. Systemic Practice and Action Research (2003) 16: 339. https://doi.org/10.1023/A:1027354823205

Al-Menayes, J. (2015). Psychometric Properties and Validation of the Arabic Social Media Addiction Scale. *Journal of Addiction,2015*, 1-6. doi:10.1155/2015/291743

American Society of Addiction Medicine, (2015), Public Policy Statement: Definition of Addiction. Retrived from https://www.asam.org/docs/default-source/public-policy-statements/1definition_of_addiction_short_4-11.pdf?sfvrsn=6e36cc2_0

Anders Koed Madsen, Mikkel Flyverbom, Martin Hilbert, Evelyn Ruppert, Big Data: Issues for an International Political Sociology of Data Practices , International Political Sociology, Volume 10, Issue 3, September 2016, Pages 275–296, https://doi.org/10.1093/ips/olw010

Andreassen, C. S. (2015). Online Social Network Site Addiction: A Comprehensive Review. Current Addiction Reports,2(2), 175-184. doi:10.1007/s40429-015-0056-9. Retrieved from https://link.springer.com/article/10.1007/s40429-015-0056-9

Andreassen, C., & Pallesen, S. (2014). Social Network Site Addiction - An Overview. *Current Pharmaceutical Design,20*(25), 4053-4061. doi:10.2174/13816128113199990616

Andreassen, Cecilie & Torsheim, Torbjørn & Brunborg, Geir & Pallesen, Ståle. (2012). Development of a Facebook Addiction Scale. Psychological reports. 110. 501-17. 10.2466/02.09.18.PR0.110.2.501-517. Retrieved from: https://www.researchgate.net/publication/225185226_Development_of_a_Facebook_Addiction_Scale

Armerding, T. (2018, December 20). The 18 biggest data breaches of the 21st century. Retrieved from https://www.csoonline.com/article/2130877/the-biggest-data-breaches-of-the-21st-century.html

Banerjee, Syagnik (Sy) and Hemphill, Thomas and Longstreet, Phil, Is IOT a Threat to Consumer Consent? The Perils of Wearable Devices' Health Data Exposure (September 18, 2017). Available at SSRN: https://ssrn.com/abstract=3038872

Beer, D. "How Should We Do the History of Big Data?" Big Data & Society, vol. 3, no. 1, 2016, p. 205395171664613., doi:10.1177/2053951716646135.

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences,116*, 69-72. doi:10.1016/j.paid.2017.04.039

Brayne, S. (2017) Big Data Surveillance: The Case of Policing. American Sociological Review, 82:5, 977–1008. https://doi.org/10.1177/0003122417725865

Burnap P, Colombo G, Amery R, Hodorog A, Scourfield J. Multi-class machine classification of suicide-related communication on Twitter. Online Soc Netw Media. 2017;2:32–44. doi:10.1016/j.osnem.2017.08.001

Cabral, J. (2011). Is Generation Y Addicted to Social Media? Elon University

Carroll, A. (1991). The Pyramid of Corporate Social Responsibility: Toward the Moral Management of Organizational Stakeholders. Business Horizons. 34. 39-48. 10.1016/0007-6813(91)90005-G. Retrieved from: https://www.researchgate.net/publication/4883660_The_Pyramid_of_Corporate_Social_Responsibility_Toward_the_Moral_Management_of_Organizational_Stakeholders

Carroll, A. (1999). Corporate social responsibility: Evolution of a definitional construct. Business & Society. 38. 268-295. Retrieved from: https://www.researchgate.net/publication/282441223_Corporate_social_responsibility_Evolution_of_a_definitional_construct

Coghlan, D. and Mary Brydon-Miller. The SAGE Encyclopedia of Action Research. SAGE Publications, 2014.

Current World Population. (n.d.). Retrieved from http://www.worldometers.info/world-population/

D. Chaffey & S. Wood. (2005). Business Information Management: Improving Performance Using Information Systems (FT Prentice Hall, Harlow).

Danah boyd & Kate Crawford (2012) CRITICAL QUESTIONS FOR BIG DATA, Information, Communication & Society, 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878

Data: Types of Data, Primary Data, Secondary Data, Solved Examples. (2018, September 21). Retrieved from https://www.toppr.com/guides/maths/statistics/data/

Davis. K. , Patterson. D.(2012) Ethics of Big Dat: Balancing risk and innovation - available at O'reilly: http://shop.oreilly.com/product/0636920021872.do

de Creek, Innovations in Clinical Practice: A Source Book, vol. 17, p. pp, Professional Definition Of Addiction. (n.d.). Retrieved from https://www.mentalhelp.net/articles/definition-of-addiction/

Denyer, David and Tranfield, David (2009) 'Producing a Systematic Review' David A. Buchanan & Alan Bryman In The Sage Handbook of Organizational Research Methods, London, SAGE, Chapter 39, Page 671-689 Detection. Retrieved from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Donkor, B. (2013, December 16). On Social Sentiment and Sentiment Analysis. Retrieved from https://brnrd.me/social-sentiment-sentiment-analysis/
Donnelly, L. (2019, January 10). Social media addicts behave like those addicted to drink and drugs . Retrieved from https://www.telegraph.co.uk/news/2019/01/10/social-media-addicts-behave-like-addicted-drink-drugs/

Dr. Young. K. , "Social Media Addiction." The Center for Internet Addiction... Your Resource since 1995, retrieved from www.netaddiction.com/ebay-addiction/.

Eijnden, R. J., Lemmens, J. S., & Valkenburg, P. M. (2016). The Social Media Disorder Scale. *Computers in Human Behavior,61*, 478-487. doi:10.1016/j.chb.2016.03.038

Eileen, F. *Action Research*. LAB, Northeast and Island Regional Education Laboratory at Brown University, 2000.

Elgan, Mike. "Social Media Addiction Is a Bigger Problem than You Think." *Computerworld*, Computerworld, 14 Dec. 2015, www.computerworld.com/article/3014439/social-media-addiction-is-a-bigger-problem-than-you-think.html?page=2.

Elphinston, Rachel & Noller, Patricia. (2011). Time to Face It! Facebook Intrusion and the Implications for Romantic Jealousy and Relationship Satisfaction. Cyberpsychology, behavior and social networking. 14. 631-5. 10.1089/cyber.2010.0318. Retrieved from: https://www.researchgate.net/publication/51104250_Time_to_Face_It_Facebook_Intrusion_and_the_Implications_for_Romantic_Jealousy_and_Relationship_Satisfaction

Flyverbom, M. & Madsen, A. (2015). Sorting data out – unpacking big data value chains and algorithmic knowledge production. Retrieved from: https://www.researchgate.net/publication/287646239_Sorting_data_out_-_unpacking_big_data_value_chains_and_algorithmic_knowledge_production

Flyverbom, M. (2017). Datafication, Transparency and Trust in the Digital Domain. In Trust at Risk: Implications for EU Policies and Institutions: Report of the Expert Group "Trust at Risk?

Foresight on the Medium-Term Implications for European Research and Innovation Policies (TRUSTFORESIGHT)" (pp. 69-84). Luxembourg: Publications Office of the European Union. DOI: 10.2777/364327

Fusch, P. I., & Ness, L. R. (2015). Are We There Yet? Data Saturation in Qualitative Research. The Qualitative Report, 20(9), 1408-1416. Retrieved from https://nsuworks.nova.edu/tqr/vol20/iss9/3

Gartner 2016, What Is Big Data? - Gartner IT Glossary - Big Data.. Retrieved from https://www.gartner.com/it-glossary/big-data/

Goodman, A. (1990). Addiction: Definition and implications. Addiction,85(11), 1403-1408. doi:10.1111/j.1360-0443.1990.tb01620.x

Günther, W. A. , et al. "Debating Big Data: A Literature Review on Realizing Value from Big Data." The Journal of Strategic Information Systems, vol. 26, no. 3, 2017, pp. 191–209., doi:10.1016/j.jsis.2017.07.003.

Günther, W. A., Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems,26*(3), 191-209. doi:10.1016/j.jsis.2017.07.003

Hu, M. & Liu, B. (2014). Opinion mining, sentiment analysis, and opinion spam
Hussain, A., and Vatrapu, R.: 'Social Data Analytics Tool', DESRIST 2014, Lecture Notes in Computer Science (LNCS), 2014, 8463, (Springer), pp. 368–372

K. S. Young, "Internet Addiction: Symptoms, Evaluation, And Treatment," in X. L. Van
Krohn, M. (2007). Robert Plutchik's Psychoevolutionary Theory of Basic Emotions.

Kruuse, A. KirkeGaard. M & Hansen. N. (2017). Machine Learning: Sentiment Analysis - Applied to Twitter and Finance - Retrieved from
https://github.com/nicklashansen/sentiment-analysis-simcorp

Kurniasih, N. (2017). Internet Addiction, Lifestyle or Mental Disorder? A Phenomenological Study on Social Media Addiction in Indonesia. *KnE Social Sciences,2*(4), 135. doi:10.18502/kss.v2i4.879

Kuss, D. J., & Griffiths, M. D. (2011). Online Social Networking and Addiction—A Review of the Psychological Literature. *International Journal of Environmental Research and Public Health,8*(9), 3528-3552. doi:10.3390/ijerph8093528
Kuss, D. J., & Griffiths, M. D. (2018, April 25). 6 questions help reveal if you're addicted to social media. Retrieved from
https://www.washingtonpost.com/news/theworldpost/wp/2018/04/25/social-media-addiction/?noredirect=on&utm_term=.44daa33887c7

Lassen, N. & Madsen, R. & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW. 2014. 81-90. 10.1109/EDOC.2014.20.

Manyika, James & Chui, Michael & Brown, Brad & Bughin, Jacques & Dobbs, Richard & Roxburgh, Charles & Hung Byers, Angela. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from: https://www.researchgate.net/publication/312596137_Big_data_The_next_frontier_for_innovation_competition_and_productivity

Mayer-Schönberger, V, & Cukier, K. (2013) Big Data: A Revolution that will transform how we live, work and think, New York: Houghton Mifflin Harcourt, Chapter 1-5.

McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. Harvard business review. 90. 60-6, 68, 128. Retrieved from: https://www.researchgate.net/publication/232279314_Big_Data_The_Management_Revolution

Meikle, J. (2012, February 03). Twitter is harder to resist than cigarettes and alcohol, study finds. Retrieved from https://www.theguardian.com/technology/2012/feb/03/twitter-resist-cigarettes-alcohol-study

Myers, S. W. (2017) Data Capitalism: Redefining the Logics of Surveillance and Privacy, Business and Society, https://doi.org/10.1177/0007650317718185

Neff, K. D. (2012). The science of self-compassion. In C. Germer & R. Siegel (Eds.), Compassion and Wisdom in Psychotherapy (pp. 79-92). New York: Guilford Press

Number of social media users worldwide 2010-2021. (2019). Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Osterwalder, A, and Y Pigneur. "Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers.", 2009. Retrieved from: www.wiley.com/en-us/Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers-p-9780470876411

Plutchik, R. (1980) Emotion: A Psychoevolutionary Synthesis. Harper & Row, New York.
Rasmussen, E.S., Østergaard, P. & Beckmann, S.C. (2006). Essentials of social science research methodology. Odense: University Press of Southern Denmark.

Rasmussen, E.S., Østergaard, P. & Beckmann, S.C. (2006). Essentials of social science research methodology. Odense: University Press of Southern Denmark.

Resource Press, Sarasota, FL, 1999.
Retrieved from: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1360-0443.2009.02828.x

Rousseau, D. M. (2006). Is there such a thing as "evidence based management"?. Academy of management review, 31(2), 256- 269.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy,Published in J. Information Science 2007, DOI:10.1177/0165551506070706

Sadock, B. J., Ruiz, P., & Sadock, V. A. (2015). *Kaplan and Sadocks synopsis of psychiatry: Behavioral sciences, clinical psychiatry*. Philadelphia: Wolters Kluwer.

Saif Mohammad & Peter Turney (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon,, In Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, June 2010, LA, California.

Salim, S. (2019). How much time do you spend on social media? Research says 142 minutes per day. Retrieved from https://www.digitalinformationworld.com/2019/01/how-much-time-do-people-spend-social-media-infographic.html

Salkind, N. J. (2010). *Encyclopedia of research design*. Los Angeles, CA: SAGE.

Sussman, Steve & N Sussman, Alan. (2011). Considering the Definition of Addiction. International journal of environmental research and public health. 8. 4025-38. 10.3390/ijerph8104025. Retrieved from: https://www.researchgate.net/publication/51786488_Considering_the_Definition_of_Addiction

Tao, Ran, et al. "Proposed Diagnostic Criteria for Internet Addiction." *The Canadian Journal of Chemical Engineering*, Wiley-Blackwell, 5 Feb. 2010, onlinelibrary.wiley.com/doi/full/10.1111/j.1360-0443.2009.02828.x.

Theory of Addiction, Second edition, West. R. , Brown. J. , 2013, Published by John Wiley & Sons.

Turel, Ofir & Serenko, Alexander. (2012). The benefits and dangers of enjoyment with social networking websites. European Journal of Information Systems. 21. 10.1057/ejis.2012.1. Retrieved from: https://www.researchgate.net/publication/263104765_The_benefits_and_dangers_of_enjoyment_with_social_networking_websites

Uckelmann, D. et al. "Architecting the Internet of Things" Springer, Springer-Verlag Berlin Heidelberg, 2011, chapter 10, www.springer.com/la/book/9783642191565.

Vatrapu, R.: 'Understanding Social Business', in Akhilesh, K.B. (Ed.): 'Emerging Dimensions of Technology Management' (Springer, 2013), pp. 147-158

Walker, L. (2018, December 24). How to Tell If You Have a Social Networking Addiction. Retrieved from https://www.lifewire.com/what-is-social-networking-addiction-2655246

Wang, Pengcheng, et al. "Social Networking Sites Addiction and Adolescent Depression: A Moderated Mediation Model of Rumination and Self-Esteem." *Personality and Individual Differences*, vol. 127, 2018, pp. 162–167., doi:10.1016/j.paid.2018.02.008.

Wilson, K, et al. "Psychological Predictors of Young Adults' Use of Social Networking Sites." *Current Neurology and Neuroscience Reports.*, U.S. National Library of Medicine, Apr. 2010, www.ncbi.nlm.nih.gov/pubmed/20528274.

World Internet Users Statistics and 2018 World Population Stats (2018). (n.d.). Retrieved from https://www.internetworldstats.com/stats.htm

Data Download Links:
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-06.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-07.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-08.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-09.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-10.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-11.txt.gz
http://snap.stanford.edu/data/bigdata/twitter7/tweets2009-12.txt.gz

9. Appendices:

[A1] Defining Addiction Links reviewed for the purpose of the Systematic Review:

http://ebookcentral.proquest.com.esc-web.lib.cbs.dk/lib/kbhnhh-ebooks/reader.action?docID=1415945 (Theory of Addiction book) Mark - Good for definition/Nichlas

https://www.mentalhelp.net/articles/definition-of-addiction/ Mark/Nichlas

https://www.addictionsandrecovery.org/what-is-addiction.htm Mark/Nichlas (Bullet points makes for a good understanding)

http://netaddiction.com/ebay-addiction/ Mark/Nichlas

https://www.lifewire.com/what-is-social-networking-addiction-2655246 Mark last two articles cover this, but the chicago experiment reference is good/Nichlas (Useful points from the uni experiment)

https://www.theguardian.com/technology/2012/feb/03/twitter-resist-cigarettes-alcohol-study Mark/Nichlas (Connected to Lifewire link and chicago experiment)

https://www.washingtonpost.com/news/theworldpost/wp/2018/04/25/social-media-addiction/?noredirect=on&utm_term=.497f2059671e Mark/Nichlas

https://smallbusiness.yahoo.com/advisor/30-signs-social-media-addiction-133619535.html Mark already covered in better articlesl/Nichlas (I feel like the article is trying to be funny/cool and not very educational/informative)

http://smaddiction.web.unc.edu/bergen-facebook-addiction-scale/ (Bergen Scale) Mark double reference/Nichlas (I do not think using double references will hurt and this one if fast and easy with the scale. Not a lot of reading and looks like a trustworthy link

http://www.thewisdompost.com/essay/addiction/social-media-addiction/social-media-addiction-meaning-symptoms-causes-effects-treatment/1293 Mark /Nichlas

https://www.telegraph.co.uk/news/2019/01/10/social-media-addicts-behave-like-addicted-drink-drugs/ Mark - SMS is as bad as other addictionss/Nichlas

https://www.itstimetologoff.com/digital-addiction/social-media-addiction-2/ Mark /Nichlas

https://www.realsimple.com/work-life/technology/social-media-addiction Mark /Nichlas

https://addictionresource.com/addiction/technology-addiction/social-media-addiction/ Mark /Nichlas

https://www.computerworld.com/article/3014439/internet/social-media-addiction-is-a-bigger-problem-than-you-think.html Mark/Nichlas

https://www.bustle.com/p/is-social-media-addiction-a-disorder-researchers-think-the-problem-needs-to-be-taken-more-seriously-15767815 Mark/Nichlas

Total:16

[A2] Decision of papers use for systematic Review:

Name	Mark	Nichlas
Addiction Protect the young from e-cigarettes.	No	No
Assessing Internet Addiction Using the Parsimonious Internet	Yes	Yes
Digital Nativity and Information Technology Addiction Age cohort versus	Partly - same as you	Partly for defining addiction type maybe
Extraversion, neuroticism, attachment style and fear of missing out as	Yes	Yes
Online Social Network Site Addiction: A Comprehensive Review	Yes	Yes
Proposed diagnostic criteria for internet addiction	Yes - we just have to remember the somewhat focus on social media when we write	Yes
Social media addiction in romantic relationships Does user's age influence	Yes - but as u mention more for the big data part	Yes, see if retweets of one person, same hash tags etc can be a proxy for addiction (need personal messages to use 100%)
Social networking sites	Yes	Yes, tells us that certain state of minds can lead to easier SNS addiction and gives basis for sentiment analysis
Emotional intelligence	Yes	Yes, same as above (only 1 page)
Addiction definition and implications	Yes	Yes, Bullet points to decide factors
BEHAVIOURAL ADDICTION OPEN DEFINITION	No	No

Considering the Definition of Addiction	Yes	Yes, use as inspiration too for a write-up
Gaming addiction, definition and measurement A large-scale empirical	Maybe - a little off the social media addiction topic? But possibly can be used as you mention	Yes, use as evidence for Internet/Behavior addicting being a thing and related to substance addiction
Public Policy Statement Definition of Addiction	Yes	Yes, addiction factors
A Quantitative Research on the Level of Social Media Addiction	Yes	Yes
Clinical Report—The Impact of Social Media on	Maybe	Partly, use for talking about social media
Comparison of factors predicting excessive	Maybe - same reasons as you	Maybe, Very Theoretical, maybe too much for systematic review
Development and validation of the Chinese social media addiction scale	Yes - mostly for sentiment analysis but could potentially be used for SR too - 6 factor.	Yes, use the scale for sentiment analysis
Exploring the role of positive metacognitions in explaining the association	No - little too much, not needed for SR i feel	Not sure, too much?
Internet Addiction, Lifestyle or Mental Disorder	Yes	Yes, Pretty good. “Addicts” Says what makes them “Addicts”
Is Generation Y Addicted to Social Media	Maybe - agree with you	Not sure, Maybe too specific/biased, else, good.
Life satisfaction A key to managing internet & social media addiction	Yes	Yes, use as a “Life satisfaction/Sentiment can tell about addiction”
Online Social Networking and Addiction—A Review of the Psychological Literature	Yes	Yes
Psychometric Properties and Validation of the Arabic Social	Yes	Yes
Social Networking Sites and Addiction	Yes	Yes

The Relations Among Social	Yes	Yes
The relationship between addictive use of social media, narcissism, and	Yes	Yes
The Social Media Disorder Scale	Yes	Yes
Time distortion when users at-risk for social media addiction engage in nonsocial media tasks	Maybe/no	Maybe
Tug of war between social self-regulation and habit Explaining the	maybe/no	Maybe for measurements

Total: 30

[A3] Dataset Description

Link	Date Accessed	Description	Main findings
https://www.mentalhelp.net/articles/definition-of-addiction/	January 08, 2019	What is addiction? Definition of Addiction	**Addiction** includes both substances and activities. **Addiction** leads to substantial harm. **Addiction** is repeated involvement despite substantial harm Addiction continues because it was, or is, pleasurable and/or valuable
https://www.addictionsanddrecovery.org/what-is-addiction.htm	January 08, 2019	What is addiction? The link gives both definition, signs, causes and consequences thereof	To be an addiction it must meet 3 or more of the following criterias: Tolerance, Withdrawal, Limited control, Negative consequences, Neglected or postponed activities, Significant time or energy spent, Desire to cut down **Causes:**

			Family history, Poor coping skills for dealing with stress, Negative thinking, Underlying anxiety or depression
http://netaddiction.com/ebay-addiction/	January 08, 2019	Social media addiction Signs, risk factors and treatment for social media addiction	**Signs:** time thinking about SNS, urge to use, use to forget other problems, restless if not using, immense time usage with negative impact on social life **Risk factors:** Teens with anxiety, depression, stress or alike are more at risk **Treatment:** Admit, Turn off notifications, schedule SNS visits, Alternate communication
https://www.lifewire.com/what-is-social-networking-addiction-2655246	January 08, 2019	What is social networking addiction?	Compulsive behaviour that leads to negative effect - Excess use of SNS Social media addiction can be stronger than addiction to cigarettes and booze
https://www.theguardian.com/technology/2012/feb/03/twitter-resist-cigarettes-alcohol-study	January 08, 2019	An experiment that explains how social media can be harder to resists than other desires - becoming an addiction easily	Social media can be more addictive/urging than cigarettes, alcohol, coffee, sex - however with less consequences, mainly “stealing time”
https://www.washingtonpost.com/news/theworldpost/wp/2018/04/25/social-media-addiction/?noredirect=on&utm_term=.497f2059671e	January 08, 2019	Review studying impact of technology and social media on cognitive and social behaviour	6 main questions to define social media addiction regarding: time usage, urge, to forget problems, restless if not using, negative impact, thoughts of reducing

			use Social media use has a significant detrimental effect on many aspects of life for a small minority - such signs are indicative of addiction similar to alcohol or drugs. Individuals are ultimately responsible for their own social media use
http://smaddiction.web.unc.edu/bergen-facebook-addiction-scale/	January 08, 2019	Scale to identify social media addiction	Social media Addiction results in several negative effects/impacts
http://www.thewisdompost.com/essay/addiction/social-media-addiction/social-media-addiction-meaning-symptoms-causes-effects-treatment/1293	January 08, 2019	Article that touch upon Signs, symptoms and causes of social media	Explains social media and its usage. Subsequently dives into symptoms and causes of addiction towards social media (mental and physical). Lists possible ways to help and cure the addiction and ends up concluding that technology is there to help society progress
https://www.telegraph.co.uk/news/2019/01/10/social-media-addicts-behave-like-addicted-drink-drugs	January 08, 2019	Research from Michigan state university on the similarities between social media and drug addiction	Social media addicts makes decisions just like substance addicted people
https://www.computerworld.com/article/3014439/internet/social-media-addiction-is-a-bigger-problem-than-you-think.html	January 08, 2019	Article that argues the social media sites are created to hook you up and make you spend more time there	FOMO and network effects in addition to software design causes addiction to social media

Research paper title (Link embedded)	Author(s) & Year	Description	Main findings

Online Social Network Site Addiction: A Comprehensive Review	Cecilie Schou Andreassen, 2015	A review of research into Social Network Site Addiction	Defines Social Network Site Addiction based on literature review. Furthermore it lists, and explored explanations, assessments, interventions and consequences of SNS addiction
Assessing Internet Addiction Using the Parsimonious Internet Addiction Components Model —A Preliminary Study	Daria J. Kuss, Gillian W. Shorter, Antonius J. van Rooij, Mark D. Griffiths & Tim M. Schoenmakers, 2013	It uses a parsimonious internet addiction components model based on 'Griffiths' addiction components to find symptoms in internet addiction associated with other types of addictions	The paper concludes that excessive use of internet can lead to various symptoms associated to other types of addictions such as gambling and online gaming, where main factors such as: salience, withdrawal, tolerance, mood, modification, relapse and conflict can measure the potential addiction.
Digital Nativity and Information Technology Addiction: Age cohort versus individual difference approaches	Hsin-Yi Wang, Leif Sigerson & Cecilia Cheng, 2018	Investigates four types of IT addiction for certain age group "digital natives": internet, smartphone, facebook and gaming disorder addictions	The study reveals a positive association between digital nativity and all the four types of IT addiction and discusses its practical implications
Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction	David Blackwell, Carrie Leaman, Rose Tramposch, Ciera Osborne, & Miriam Liss, 2017	The paper studies whether extraversion, neuroticism, attachment style and FOMO could be used as predictors for social media use and prediction, using a survey for 207 participants.	Using methods such as bergen scale, big five inventory and FOMO scale to collect relevant data, it was deemed that extraversion, neuroticism, attachment style and FOMO were all significant to use as predictors for social media use and addiction. Especially FOMO.

Proposed diagnostic criteria for internet addiction	Ran Tao, Xiuqin Huang , Jinan Wang, Huimin Zhang, Ying Zhang & Mengchen Li, 2009	Studies a way of diagnosing internet addiction	It was found that a presence of 2 symptoms (preoccupation and withdrawal symptoms) accompanied by at least 1 other of five symptombs (tolerance, lack of control, continued excessive use despite knowledge of negative effects, loss of interests excluding internet and use of internet to escape issues) could be used to accurately diagnose internet addiction
Social media addiction in romantic relationships: Does user's age influence vulnerability to social media infidelity?	Irum Saeed Abbasi, 2018	The paper examines SNS addiction in relation to romantic relationships / affairs and whether or not age influences this.	SNS can predict infidelity related behaviors and age moderates the connection. Higher SNS addiction promotes infidelity behaviour being more present at lower ages
Social networking sites addiction and adolescent depression: A moderated mediation model of rumination and self-esteem	Pengcheng Wanga, Xinyue Wanga, Yingqiu Wub, Xiaochun Xiec, Xingchao Wanga, Fengqing Zhaoa, Mingkun Ouyanga & Li Leia, 2018	Assesses the connection between SNS and depression. It does so by examining whether rumination mediated the relation between SNS addiction and depression and whether the mediating effect was moderated by self esteem	SNS addiction is positively associated with depression, where rumination mediated this relation. Furthermore lower self esteem is stronger related to this than high self esteem. The SNS addiction was measured with the facebook intrusion questionnaire.
The Relationship between Emotional Intelligence and Technology Addiction among University	Jalaleddin Hamissi, Mohadese Babaie, Mehdi Hosseini & Fatemeh Babaie, 2013	Study that examines the connection between internet addiction, virtual environments and	High Emotional intelligence and the severity of internet addiction has an inverse relationship.

Students		emotional intelligence amongst university students	It also shows the results of internet addiction amongst 201 university students
Addiction: definition and implications. British Journal of Addiction	Aviel Goodman, 1990	A journal commentary on how the lack of a scientific Definition of addiction hampers the integration of addiction, both in theory and practise	Proposes a definition of addiction grounded in the DSM-III-R. The definition is formulated in general terms, hence not restricted by reference to particular behaviour
Considering the Definition of Addiction	Steve Sussman & Alan N. Sussman, 2011	A review of 52 studies that explores the definition of addiction, deriving A through Elements and researching differences from compulsion	Engagement in the behavior to achieve appetitive effects, preoccupation with the behavior, temporary satiation, loss of control, and suffering negative consequences are all elements of addiction that leads to negative consequences. What is addiction and not “When can you be defined as an addict”

Public Policy Statement Definition of Addiction	American Society of Addiction Medicine, 2011	Public Policy Statement from the American Association of Addiction Medicine (ASAM)	Addiction is a primary, chronic disease of brain reward, motivation, memory and related circuitry. Addiction is characterized by inability to consistently abstain, impairment in behavioral control, craving, diminished recognition of significant problems with one's behaviors and interpersonal relationships, and a dysfunctional emotional response.
A Quantitative Research on the Level of Social Media Addiction among young people in Turkey	Ali Murat Kirik , Ahmet Arslan, Ahmet Çetinkaya & Mehmet Gül, 2015	Research of young Turkish people and their relation to social media	Touches upon social media and what it is. Furthermore it explains the connection that people seeks on social media. This Article will NOT be used for its specific turkish research but for its explanatory value
Development and validation of the Chinese social media addiction scale	Chang Liu & Jianling Ma, 2018	Defining a specific social media addiction scale for the chinese - what is important to consider to assess social media addiction (experimented on 619 college students)	A six-factor model was deemed useful for assessing social media addiction. This model takes the following factors into account: social interaction, mood alteration, negative consequences and continued use, compulsive and

			withdrawal factor, salience and relapse. Social media addiction is also correlated with smartphone addiction, pathological internet use and narcissism, but negatively associated with self-esteem.
Internet Addiction, Lifestyle or Mental Disorder? A phenomenological study on social media addiction in Indonesia	Nuning Kurniasih, 2017	A phenomenological (in depth) study on social media addiction in Indonesia - it focuses on main reasons for internet addiction, habit of usage, time use, feeling when accessing, feeling when not able to access, correlation between internet addition and working performance, addictive points about internet and how to distract informants from the internet	By focusing on four significant points regarding internet addiction: excessive use, withdrawal, tolerance and negative repercussions. It was possible to asses the informants ("addicted people") regarding their social media addiction, ultimately concluding: up to 8 hours use daily, upset when disconnected, negative repercussions especially regarding family and more. Moreover, it was stated that subjects can be distracted from social media with substituted behaviour that can give this certain pleasure that SNS does, such as hobbies, social interaction travelling etc.. .

Life satisfaction: A key to managing internet & social media addiction	Phil Longstreet & Stoney Brooks, 2017	The Role of life satisfaction(Happiness, stress etc.) as a tool to reduce generalized and internet/social media addiction	The study estimates 444 Furthermore, results shows that the life satisfaction of an individual effects addiction behaviour, and that this addiction might be deeply rooted in different issues in their lives, and that these issues lowers their life satisfaction and increases vulnerability towards addiction.
Online Social Networking and Addiction—A Review of the Psychological Literature	Mark D Griffiths & Daria Kuss, 2011	Outlines SNS usage patterns, examines motivations for SNSs usage, personalities of SNS users, negative consequences of SNS usage, explores potential SNS addiction and explores SNS addiction specificity and comorbidity	SNS is predominantly used for social purposes Extraverts appear to use SNS for social enhancement and introverts for social compensation each of which are related to greater usage along with low conscientiousness and high narcissism Negative correlates of SNS usage are: decrease in real life social community participation and academic achievement as well as relationship problems indicating potential addiction
Psychometric Properties and Validation of the Arabic Social Media addiction scale	Jamal Al-Menayes, 2015	Studies psychometric properties using social media addiction scale - a variant of Internet addiction test	Two most important factors to measure SMA are: *dependent use and excessive use*

			The scale uses: salience, excessive use, neglect of work, anticipation, self-control and neglect of social life
Social Networking Sites and Addiction	Mark D Griffiths & Daria Kuss, 2017	New insights into SNS and addiction providing 10 lessons learned from other literature	**Social networking** and social media use are not the same **Social networking** is eclectic **Social networking** is a way of being. **Individuals** can become addicted to using social networking sites **Facebook** addiction is only one example of SNS addiction **Fear of missing** out (FOMO) may be part of SNS addiction **Smartphone addiction** may be part of SNS addiction **Nomophobia** may be part of SNS addiction There are **sociodemographic differences** in SNS addiction There are **methodological problems** within the research.
The Relations Among Social Media Addiction, Self-Esteem, and Life Satisfaction in University Students	Nazir S. Hawi1 & Maya Samaha1, 2017	How social media addiction is connected to self-esteem and life satisfaction for university students, using SMAG Rosenberg's self-esteem scale and satisfaction with life scale	One-factor model of SMAG has good psychometric propertie and had high internal consistency. Addiction to social media had negative association with self-esteem, as people with lower self-esteem scored higher on social media addiction. Self-esteem

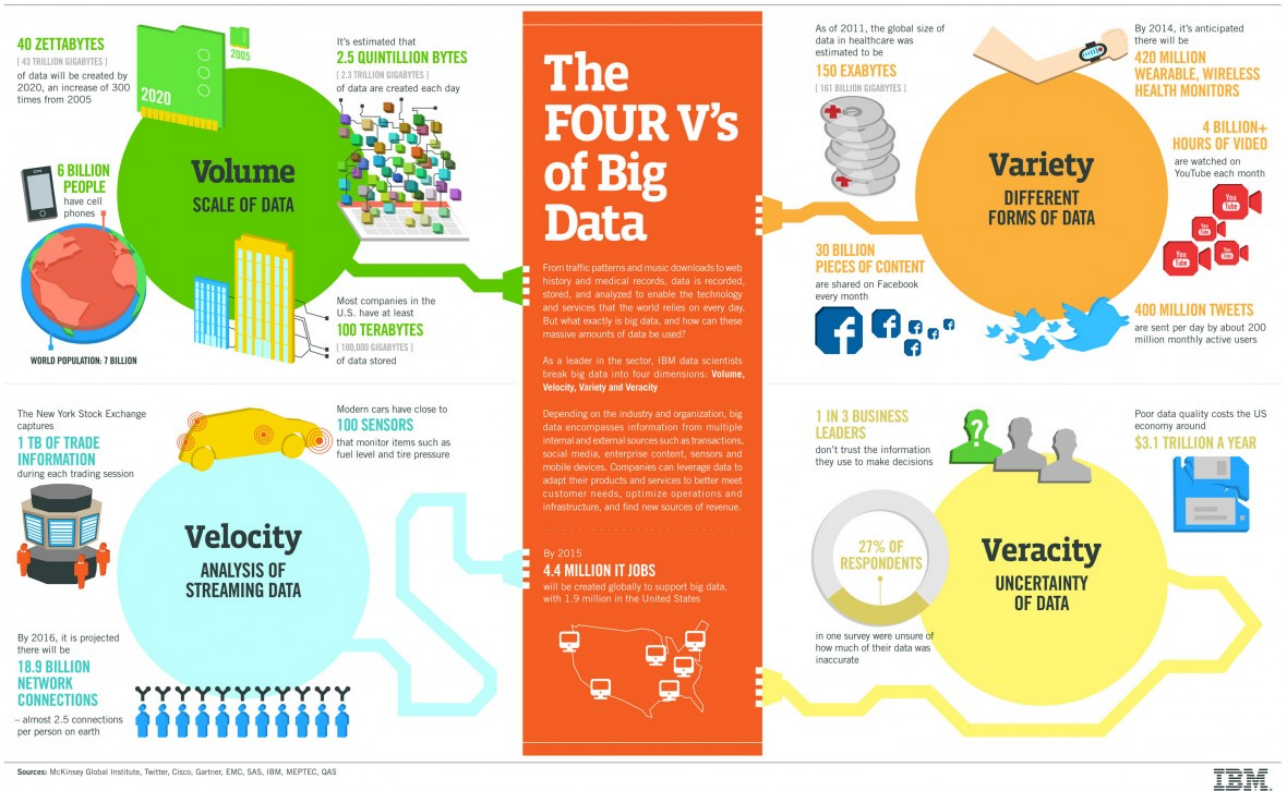
			had a positive association with satisfaction with life. Furthermore there is a mediated negative relationship between social media addiction and satisfaction with life.
The relationship between addictive use of social media, narcissism, and self-esteem	Cecilie Schou Andreassen, Ståle Pallesen & Mark D. Griffiths, 2016	Large national survey that uses bergen scale, narcissistic personality inventory-16 and rosenberg self-esteem scale to measure social media addictiveness and connecting it to narcissism and self-esteem	Lower ages, women in no relationship, students with low income, lower self esteem and narcissism are associated with social media addiction
The Social Media Disorder Scale	Regina J.J.M. van den Eijnden, Jeroen S. Lemmens & Patti M. Valkenburg, 2016	Creates a social media disorder scale to create a clear diagnostic tool to distinguish between social media disorder and high-engaging non-disordered social media users	A 9-item scale defined as valid and psychometrically sound to measure social media disorder using the following items: Preoccupation, tolerance, withdrawal, displacement, escape, problems, deception, displacement and conflict Furthermore found that looking at self-esteem, depression, attention deficit, loneliness and frequency of daily social media use is useful for diagnosing and related to the 9 item scale.

[A4] The bergen Scale (Andreassen et al, 2012)
Six-question questionnaire:

Answer:
1. Very rarely
2. Rarely
3. Sometimes
4. Often
5. Very often

To:
1. You spend a lot of time thinking about Facebook or planning how to use it.(Salience)
2. You feel an urge to use Facebook more and more(Tolerance).
3. You use Facebook in order to forget about personal problems(mood modification).
4. You have tried to cut down on the use of Facebook without success(Relapse).
5. You become restless or troubled if you are prohibited from using Facebook(Withdrawal).
6. You use Facebook so much that it has had a negative impact on your job/studies(conflict).

[A5] IBM's Four V's of big data



[A6] Robert Plutchik's PSYCHOEVOLUTIONARY THEORY OF BASIC EMOTIONS