Social Media and Forecasting: What is the Predictive Power of Social Media?

Leah J. Schade
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

The emergence and growing usage of Social Media has resulted in extremely vast amounts of data that can be collected and accumulated. Businesses are furthermore increasingly interested in using customer generated Social Media Data to generate useful insights about their customers. However, the data from Social Media platforms is very complex in nature and thus poses a variety of challenges which have to be overcome to successfully implement Social Media based forecasting.

The purpose of this paper is consequently to evaluate upon Data processing techniques that can be used to make predictions about future customer needs and future market trends. A critical literature review has discussed and critically elaborated on past findings in Social Media based forecasting, presented different data processing techniques, highlighted the importance of organizational prerequisites and pointed out limitations of making predictions based on data derived from Social Media activity.

The results indicate that Social Media data can indeed be used to make predictions about market needs and market trends if analyzed in a professional manner. The paper additionally presents a broad overview of tools that can be used for the areas of data collection, database management and data analysis. Lastly, in an attempt to position this paper as helpful for practitioners a framework is proposed to enable effective forecasting with Social Media Data.

Supervisors: Dr. E. Constantinides & J. van Benthem

Keywords

Social Media, Big Data, Data Processing, Data Mining, Social Media Analytics, Forecasting, Customer Needs Market trends

5th *IBA Bachelor Thesis Conference*, July 2nd, 2015, Enschede, The Netherlands.

1. INTRODUCTION

1.1 Background Information

Since the advent of Social Media, the World Wide Web has progressed from a state where users used to passively consume information that is provided to them for example by marketers to a diverse Social Media landscape in which users can constantly and actively share and update information (Xu & Liu, 2014). Kietzmann, Hermkens, McCarthy and Silvestre (2011) imply that "Social Media employ mobile and web-based technologies to create highly interactive platforms via which individuals communicate" (p. 241). Cooke and Buckley (2008) characterize a society that is increasingly willing to share their experiences, opinions and thoughts on the so-called "Social Web" and contrary to the traditional media, users of Social Media can publish their own ideas and information online for everyone to read (Hildreth & Ament, 2011; Hoffman & Fodor, 2010). The emergence of a social media ecosystem and the growing usage of social media platforms such as Facebook, Twitter, YouTube and Instagram, only to name a few of the largest ones, have led to a tremendous amount of user-generated content, covering a large variety of topics. Social Web is another term, next to Social Media, that is used to describe the multifaceted Social Media landscape which essentially consists of forums, discussion boards, podcasts, social networks, blogs, microblogs, online chats, video sharing planforms, and photo sharing sites which are originally created to facilitate a social exchange and interaction between people and organizations (Hildreth & Ament, 2011). The Social Media expansion and the information that is shared by millions of users on Social Media platforms create a massive amount of easily accessible data about individuals, society and online behavior of consumers (Schoen et al, 2013). This massive amount of collected data which is referred to as Big Data opens up a whole new field of research for statisticians, social scientists, economists and moreover yields a great opportunity for making predictions about the future (Schoen et al., 2013). In fact, quite some research about the predictive qualities of social media data has already been conducted in the areas of politics, economics, and social sciences and in the health care sector. Bollen, Mao and Zeng (2010) for example analyzed the mood of Twitter posts in order to make predictions about the stock market, Gayo-Avello (2012) investigated in the direction of predicting election outcomes using Twitter data and Polgreen et al. (2008) deliberated on the predictability of an influenza outbreak through monitoring influenza-related internet searches. However, while this early literature provides essentially important insights, not much research has been done in the area of customer-generated data as a means of creating business value in terms of forecasting market trends, customer needs and customer behavior which represents a research gap.

1.2 Research Problem

With an increased Social Media and Internet usage, most people cannot imagine life without Social Media anymore as it facilitates an easy way of communicating, staying in touch and sharing moments of their lives in form of audio, pictures, videos and text with friends and family. Therefore people spend more and more time being active on the Internet and of course on Social Media platforms. Especially since the costs of storing vast amounts of data are decreasing , every mouse click and action of users on the internet is stored and leaves data traces of customer online behavior including what they search for, what they buy, what they watch, what they talk about and what they share (Jungherr & Jürgens, 2013). In today's diverse social media ecosystem firms can easily find and accumulate tremendous amounts of customer data from market research, customer interactions and transactions on Social Media but also from various types of mediums and sensors including for example cameras, GPS, RFID chips and accelerometers. Fania and Miller (2012) describe that the exponentially growing amount of raw data is next to social media, sensors and cameras additionally a result of the explosion of connected devices. The so-called 'Internet of Things' (IoT), "connects devices such as everyday consumer objects and industrial equipment onto the network, enabling information gathering and management of these devices via software to increase efficiency, enable new services, or […] other environmental benefits" (Jankowski, Covello, Bellini, Ritchie & Costa, 2012, p.2). Ever more data is generated online by the minute with firms being increasingly interested in hoarding and analyzing customer data in order to create business value. Constantly growing and changing data, various types of data, the limited amount of experienced professionals in the field of data processing and the need for speedy analysis techniques to be able to provide timely and significant data shape the major problem for firms (Fania & Miller, 2012). Therefore, the big question for firms is how they can make sense of the immense amount of data collected from social media in order to produce useful insights into future customer behavior, future customer needs and into upcoming market trends. Many firms are either unaware of the possibility of gaining insights about their customers' needs and future market developments from the large data sets that have been generated through their customer's social media activity, are unsure of how they should pursue this approach or lack the right technology and knowledge to effectively utilize Social Media Big Data (American Marketing Association, 2012). When looking through the literature it becomes apparent that not much research has been conducted in the field of predicting market trends or customer needs. This paper will try to fill this gap by providing insights about the predictive potential of Social Media data in regards to future market needs and developments. Derived from the previous introduction to the topic and its defined research problem, the following research question evolved: *What is the value and predictive potential of customer generated Social Media data for firms in order to create business intelligence concerning customer needs, innovation and market trends?* On the basis of this research question the following sub-questions will be addressed within the literature review of this paper:

1.2.1 Sub-Questions

- *Based on the existing field of literature, what has been predicted so far using Social Media data?*
- *How can Social Media data be used and processed in order to extract meaningful information?*
- *What organizational prerequisites are needed to extract a predictive potential from Social Media data in order to create Business Intelligence?*
- *What are the limitations of Social Media based forecasting?*

The remainder of the paper is structured as follows: The introduction of the topic is followed by a literature review that presents the existing academic literature in the field of forecasting using Social Media data. The literature review will be fragmented into 4 sections in which each of the sub-questions will be addressed respectively. Section 2.1 presents the diverse field of literature in the area of forecasting with data derived from social media usage. The following section (2.2) discusses Data Processing Techniques including Data mining, and Social Media Analytics. In addition, an overview of organizational prerequisites will be presented, followed by limitations and criticism of Social Media data as a forecasting tool. The findings of the literature review will be critically commented on and discussed in section 3, which is then followed by a framework as to how Social Media can be used to predict future needs of market trends. Lastly, the part 'Discussion' will include a summary of the most important findings, limitations of the paper and suggestions for further research and an explanation of the relevance of this paper.

1.3 Methodology

A critical literature review comprises an overview of published scientific material in order to provide an impression of the current state of knowledge in the field in question but also comprises of using the existing knowledge and to add value to that knowledge by comparing it and critically commenting on it. Based on the Guidelines for writing a literature review from Mongan-Rallis (2014) [1] this paper critically reviews and systematically analyzes 57 relevant scientific papers selected from the existing literature in order to identify the forecasting potential of Social Media Data. The criterion for selecting the chosen papers was based upon a focus on Social Media Big Data. A lot of articles cover the predictive potential of Big Data in general without discussing the aspect of Big Data deriving from customer activity in Social Media in detail. Academic papers for this critical literature review were mainly found using online search engines for finding scientific articles such as Scopus, Google Scholar or the University of Twente online library. The following keywords 'big data' 'data mining' 'data processing' 'analytics' together with the catchwords 'social media' 'forecasting' 'prediction' 'predictive' were used on those aforementioned electronic search engines to find suitable literature. On Scopus, the literature that was found using the key search terms was first filtered according to year of publication in order to find very recent academic works since the topic of Social Media based forecasting is still fairly new, and second it was sorted according to the amount of citations in order to find relevant and much-discussed articles. Furthermore, 'snowballing' or 'reference harvesting' was carried out in order to find additional and frequently cited literature. After the initial discovering phase, the collected literature was thoroughly analyzed in order to eliminate those articles that proved to be not necessarily relevant to the topic of this critical review. This in depth analysis begins with a read through the abstract of each paper, followed by a closer look at the introduction and the conclusion. These three parts already convey some of the main insights of the article under study and help to understand in what way the article is relevant to the topic. If the article still seems to be of importance after the forgone steps, it will be studied as a whole. The remaining articles were then divided into sub-categories that reflect the topics of the sub-questions that were posed in the description of the research problem.

Social Media based prediction is a very young topic which is why much of the literature used in this paper are proceedings of conferences. Furthermore, due to the fact that Social Media and in particular the predictive power of Social Media is a fairly new field of research, all papers and conference proceedings which are examined in this paper are published after the year of 2001 apart from for one exception.

2. LITERATURE REVIEW

2.1 Social Media based Forecasting – Former Research

Since the advent of Social Media, quite an extensive amount of research has been conducted trying to discover the topic of social media as an enabler of forecasting what is likely to happen in the future. Several authors have created reasonably successful outcomes with their research efforts, especially in the fields of epidemiology, economics, politics and social science. Early research about predicting with data derived from the Internet was carried out within the health care sector. Polgreen et al. (2008) and Ginsberg et al. (2009) have both used search engine query data to detect and predict influenza outbreaks ahead of time. Corley, Cook, Mikler and Singh (2010) have furthermore evaluated text and data mining on the web and social media to monitor influenza trends. Achrekar, Ghande, Lazarus, Yu and Liu (2011) and Signorini, Segre and Polgreen (2011) claim to be able to predict flu trends and track levels of disease activity with the help of their own Flu Trend framework and by analyzing the public sentiment of Tweets, respectively. Sakari, Okazaki and Matsuo (2010) and Earle, Bowden and Guy (2011) have created models that detect the occurrence of earthquakes by means of analyzing tweets based on context and location estimation to forecast the further expansion of earthquakes. Another segment of research that has attracted much attention is dedicated to predicting stock market developments. Bollen et al. (2010) analyzed and assessed changes in the public mood of large-scale collections of Twitter posts in order to make predictions about future movements on the stock market. Similarly, Gilbert and Karahalios (2009) and Zhang, Fuehres and Gloor (2011) assessed the public mood and presented findings that an increased expression of anxiety and emotional outbreaks can predict negative pressure on Dow Jones, NASDAQ and S&P 500. Yet another very commonly, if not the most commonly investigated topic is the issue of predicting election outcomes using Social Media data (Franch, 2013; Metaxas, Mustafaraj and Gayo-Avello, 2011; Boutet, Kim and Yoneki, (2012) & Jahanbakhsh and Moon, 2014). Especially Twitter content has frequently been examined as a basis for making predictions about elections. Tumasjan, Sprenger, Sadner and Welpe (2010) for example claimed that it would be possible to predict election outcomes by observing the frequency of mentions of political parties during the election phase on Twitter and Jungherr, Jürgen and Schoen (2011) replicated this empirical study and concluded that the amount of mentions of political parties on Twitter is not effective as a predictive indicator for election outcomes. Asur and Huberman (2010) have created a simple model that can forecast box-office revenues of movies based on the rate at which Tweets are created about a certain topic on Twitter. They also show that the forecasting ability of this model can even further be improved by including an analysis of the sentiment of Tweets. Another empirical study carried out by Mishne and Glance (2006) analyzed weblog content in particular sentiment to predict movie success and found a positive correlation between blogger sentiment and movie success, however did also admit that the correlation was not high enough to build a predictive model based on sentiment alone. Other research efforts have aimed at predicting product sales, relationship tie strength, and individual behavior using Social Media Data (Ghose & Ipeirotis 2011, Gilbert & Karaholios, 2009a, Goel & Goldstein, 2014). The presented literature shows that the approaches of predicting

[1] Retrieved from: http://www.duluth.umn.edu/~hrallis/guides/researching/litreview.html

with Social Media Data are targeting very different areas of interest and that a number of different approaches were used to investigate the predictive potential of Social Media. Most of the findings in the topic of forecasting with Social Media data are claiming to have produced predictive outcomes to some extent; however there are also controversial research outcomes as described above in the case of Jungherr et al. (2011) and Tumasjan et al. (2010).

2.2 Data Processing in the era of Big Data

The term ‘Big Data’ is commonly used to describe such immensely large sets of data that require the use of special machinery in order to manage them. Russom (2011) further explains that data volume (or quantity of data) is however not the only defining attribute of Big Data and that next to the data volume two other characteristics are of great importance, building the three V’s of Big Data. These other two characteristics are data variety, the nature of the data type (structured, unstructured and semi-structured data) and data velocity, desired analysis rate (batch, near time, real time and streams) (Russom 2011). Complex, unstructured or semi-structured data is data derived from text, images, audio and/or video and describes data in which the connections between the data elements are not clear and that probabilities will have to be determined (Bloem, van Doorn, Duivestein, van Manen and van Ommeren, 2012). Bloem et al. (2012) further claim however that the difficulties arising from having structured, unstructured and/or semi-structured data are decreasing over time as technology and IT solutions for data collection and data processing are not only prospering but are getting increasingly more affordable which ultimately means that a growing number of firms will be able to apply data management techniques in order to create competitive advantage. Similarly, Boyd and Crawford (2012) also state that Big Data is not about the size of the data set but rather define it through an interplay of technology (maximizing computation power and algorithmic accuracy), analysis (drawing on large data sets to identify patterns) and mythology (the widespread belief that large data sets offer a higher form of intelligence) (p.663). Bloem et al. (2012) further explain that “the rules of the game remain the same, but the tactics have changed” (p.8) which essentially means that the fundamental process of using information from raw data to create business insights and to enable better decision making is similar in traditional ways and in the era of social media yet the methods of tackling the processing step are changing. However, Barbier and Liu (2011) highlight that there are three characteristics of Social Media Big Data that challenge researchers and the traditional ways of mining Big Data, that is that social media data is large, noisy and dynamic. The characteristic ‘large’ refers to the excessive amount of data that is created by millions of users and Social Media data is in so far ‘noisy’ that not all data from Social Media is relevant and includes for example spam blogs which creates a load of data that has to be filtered out in order to ensure that the data is meaningful for the target application (Barbier & Liu, 2011). The last characteristic described by Barbier and Liu is ‘dynamic’ and essentially describes the fast changing nature of the online social media world. In this sense, data processing describes the process of collecting, arranging and analyzing Big Data in a way that produces significant information that can be used to support better and faster business decisions. Even though “Big Data is now becoming recognized as the newest strategic asset, a goldmine for actionable insight into every aspect of one’s business” (Fania & Miller, 2012, p.2), it is nonetheless important to emphasize that Big Data from Social media poses a variety of challenges, in particular its size, noisiness, dynamic nature and variety that need to be overcome in order to successfully analyze data sets and to extract useful and significant information and business insights from them.

2.2.1 Data Processing Techniques

During the mid-2000’s, in the beginning stage of Social Media firms would simply have to watch out for customer postings on the firm’s own website to figure out if customers might be unsatisfied. Firms could then work towards resolving the issue at hand (Fan & Gordon 2014). About a decade later, the situation has changed drastically with millions of users posting about firms on various Social Media platforms. For firms this essentially means that monitoring by itself is not sufficient anymore. And as previously described, firms are drowning in data especially since the emergence of Social Media that brings along ever more data that can be accumulated and analyzed. The amount of raw data that firms collect however has no power if it is not handled and analyzed professionally in order to discover meaningful and predictive insights. The enormous amount of data that is amassed is becoming far too large to be analyzed manually by humans alone, which reveals the need for automated technology and tools that can assist in extracting knowledge from Social Media Data. Barbier and Liu (2011) further claim that the size of Social Media data sets makes it almost impossible to analyze them without the use of automated information processing techniques and social media analytics. This is where data processing techniques come into place, offering a way of making sense of the vast amount of Social Media data every day and if carried out in a professional manner, data processing techniques can create a significant competitive advantage and revenue growth. Based on the three V’s of Big Data, Volume, Velocity and Variation, traditional methods that were used to manage, analyze and process data are not fitting in a Big Data environment as they are unable to handle the complexity of Big Data (American Marketing Association, 2012). The following sections will therefore present a taxonomy of the most commonly used data processing techniques.

2.2.1.1 Data Mining

This section presents the key purpose of data mining, a description of the preceding technological developments, and an overview of the methods used to gain useful insights.

Data Mining is essentially an analytical process intended to handle and analyze large amounts of data in order to identify underlying patterns, trends and correlations, that are not readily apparent, and which then can be used to provide useful insights about real-life problems (Barbier & Liu, 2011). The crucial idea behind data mining is therefore to find new information in a data set and to better understand large sets of data (Barbier & Liu, 2011). The Social Media environment of today offers a completely new kind of customer data that yields information and insights about communication, society and social networks in general (Barbier & Liu, 2011). The process itself involves the application of varying mining techniques in order to reveal an otherwise hidden pattern. While some regard Data mining as one of seven steps in knowledge discovery from data (short KDD) others, especially in the research milieu classify all steps of knowledge discovery from data, which are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation under the overall term Data mining (Fayyad, Piatetsky-Shapiro & Smyth, 1996; Han, Kamber & Pei, 2012).

According to Han et al., 2012, data mining is a natural and logical sequel to the previous evolution of information technology. The development in the information technology industry can be briefly depicted as follows: Starting with the premature data collection and database creation as a foundation for the later developments in the 1960's, to more advanced and sophisticated data management systems in the 1970's. From then on the development moved towards advanced database systems including data warehousing and data mining to allow for advanced data analysis (mid-1980's until present). Based on the previous progress especially in the supply of powerful and affordable computers, data collection equipment and storage media, in the late 1980's advanced data analysis was the next huge step and is still on-going. From this short overview of the advancement in database and data management it becomes clear that data mining is not a novel practice in the field of data processing in fact, business, scientists and governments have all used data mining for quite some time to convert raw data into predictive insights. Data mining is specifically valuable to data processing as it is applicable to a vast variety of data as long as the data is significant for the target application, there are however critics who have claimed that one can find underlying patterns in any kind of data, that appear to be significant even if they are not, if they investigate the data long enough (Fayyad et al. 1996 & Han et al., 2012). Data mining methods are mainly established on techniques from machine-learning, pattern recognition and statistics and consist of an iterative application of different fundamental methods such as classification, statistical analysis, and clustering (Fayyad et al. 1996). Classification can be described as the task of a function that classifies an item of data into a predefined class or under a predefined categorical label such as "yes" or "no". The function is derived from a training dataset for which labels are known and is then used on a dataset where labels are not known to predict class labels (Han et al, 2012). Statistical models are often used to make sense of the data at hand and are a fairly prevalent method of analyzing customer data from Social Media activity in order to make predictions about outcomes (Schoen et al., 2013). Regression analysis is a statistical method that is rather used for numeric data processing. It is explained by Fayyad et al. (1996) as "learning a function that maps a data item to a real-valued prediction variable" (p. 44). Linear regression models analyze the relationship between a dependent and one or more independent variables and are the simplest and most commonly used (Yu & Kak, 2012). Data mining is however not to be confused with statistics as statistics primarily focuses on organizing and summarizing information whereas data mining is primarily concerned with finding hidden patterns within the data sets (Han et al., 2012). Clustering or a cluster analysis can be used to group together similar objects of data that have no predefined label in order to generate an overall label for each cluster. The resulting clusters can vary according to the different clustering algorithms that can be used and is very helpful in identifying previously unknown groups within the data (Han et al. 2012).

Section 2.2 laid out the main challenges in analyzing and handling Social Media Data and it has to be pointed out that other data sets may contain some of the same difficulties as mentioned in section 2.2 but typically not all of them at once as in the case of Social Media Data (Barbier and Liu, 2011).Barton and Court (2012) further criticize that simple data mining runs numerous statistical tests to detect underlying patterns, yet leaves practitioners trying to figure out what the data really says in terms of business performance. In conclusion it can be said that Data mining provides the basis for analyzing large sets of data but are not as successful in handling the amount of data that is generated by Social Media all at once. In order to effectively investigate Social Media Big Data at once, more recent developments in data processing approaches might be more appropriate. The following section will therefore introduce the area of Social Media Analytics.

2.2.1.2 Social Media Analytics

This section provides definitions of the key terms and focuses especially on the goal of Social Media Analytics, the main processing and analysis methods of Social Media Data and introduces some tools that can be used for the implementation for the each of the methods. When browsing through recent literature one often comes across the terms Big Data Analytics and Social Media Analytics. The terms are repeatedly used interchangeably and in essence describe the process of measuring, analyzing and interpreting Social Media Data. Gandomi and Haider (2014) define Social Media Analytics as an analysis of the structured and unstructured data that is generated on online Social Media platforms and mention its data-centric nature as a main characteristic. In this sense the term "Predictive Analytics" is also recurrently mentioned. One definition of the term predictive analytics is from Shmueli and Koppius (2011) and states that "predictive analytics include statistical models and other empirical methods that are aimed at creating empirical predictions, as well as methods for assessing the quality of those predictions in practice" (p.554). Even though statistical methods are used, predictive analytics in in essence different from statistics in that it is quantitative and qualitative rather than simply quantitative (Waller & Fawcett, 2013). Fan and Gordon (2014) describe the process of Social Media analytics as a three stage process which involves capturing, understanding and presenting data. Data is captured by monitoring and 'listening' to a wide range of social media platforms in order to extract information that might be of relevance for the company's interest. Data is collected by going through masses of social media content using for example news feeds, application programming interfaces or web crawlers that systematically search the World Wide Web. The second step in the process, named 'understand', is the core stage in the process and will provide the input and basis for the third stage. Within this stage, several valuable metrics and trends can be produced that are able to provide insights about customer backgrounds, interests, concerns and networks of relationships. The enormous amount of raw data will be cleared out from irrelevant and noisy low-quality data with the help of text classifiers so that only pertinent information is left for the main analytical part in which meaning is derived from the data using statistical and analytical techniques. Within the third stage results from the prior stage are being summarized, evaluated and presented in a meaningful and easily understandable way. It is important to note that in order to keep up to date with possibly fast changing perceptions in the market; the process needs to be viewed as ongoing and iterative. As previously mentioned, Social Media analytics entail various techniques from quite different fields. Some of these techniques will be described in the remainder of this paragraph.

The topic of sentiment analysis or opinion mining, as it is often also referred to, has generated a lot of interest among researchers especially since the rise of Social Media and thus the availability of datasets which can be used to train machine learning algorithms as well as the development of machine learning methods in natural language processing and information retrieval (Pang & Lee, 2008). It has been suggested that the sentiment of people towards a specific topic can provide very valuable information about for example brand perception, new product perception and overall firm reputation (Barbier & Liu, 2011; Gundecha & Liu, 2012). Within a sentiment analysis computational linguistics, natural language processing and text analytics are used to acquire user sentiment and opinions from Social Media user-generated data (Fan & Gordon, 2014). A very traditional and simple method within a sentiment analysis is word count. It is generally assumed that the more often a word or brand name is mentioned, the more it can be assumed that the brand or product is liked (Fan & Gordon, 2014). However, contrarily to this belief is the statement of Mejova (2009) who claims that "in the field of Sentiment Analysis we find that instead of paying attention to most frequent terms, it is more beneficial to seek out the most unique ones" (p.9). This belief is based on the study of Wiebe, Wilson, Bruce, Bell and Martin (2004) in which they discover that people can be quite creative when they are opinionated thus using unique words to explain their position. A second approach is the lexicon approach. It includes very simple polarity lexicons with binary classifications of words into for instance "good" and "bad" or "positive" and "negative" (Fan & Gordon, 2014 & Mejova, 2009). This simple method can act as the basic for finer distinctions such as using semantic methods for computing lexical distances between a product and each of the binary terms to determine sentiment or defining the strength of affect level (intensity) and degree of relatedness to the category (centrality) of certain words (Fan & Gordon, 2014 & Subasic & Huettner, 2001). One of the most popular lexicons that has been used for sentiment analysis is WordNet (Mejova, 2009). WordNet which has been created at the Princeton University is a lexical database for the English language that can be used for automated text analysis for example in the field of a sentiment analysis (Miller, 1995; Mejova, 2009). Other tools that can be very helpful are for example ethority Gridmaster, a tool that analyzes Social Media data based on sentiment analysis and Foodmood.in. FoodMood.in is based on the Stanford University sentiment classifier, that can handle millions of Tweets, and combines Tweets with locations in order to measure the overall current mood in regards to food in a certain region. The tool is also able to portray individuals and their specific experiences and complete Tweets about food next to the overall sentiment in the location or region. (Bloem et al., 2013) For an extensive and thorough state of the art sentiment analysis the author would like to refer to Pang & Lee (2008) and Liu (2012). One downside of sentiment analysis is however the fact that it can produce false information based on incorrect readings that are attributable to the complexity of human conversation including for example humor and sarcasm which are not easily detectable using machine learning or lexicon approaches (Metaxas et al., 2011).

The next approach that is presented is a social network analysis. Social Networking sites like Facebook, Xing or LinkedIn consist of connected users that have their own profile and share information in the form of for example news posts, photos or videos with their 'friends' (Barbier & Liu, 2011). Social Network analysis views social relationships in terms of nodes, representing the individual actors (or users) within a network and of ties (or often called links) standing for the relationships between the individual actors in the network (Yu & Kak, 2012). Nodes can furthermore represent customers, end-users, and products or services and nodes may also symbolize collaboration, transactions and/or email exchange depending on the area of interest (Chen, Chiang & Storey 2012). The main goal of a social network analysis is to understand the underlying structure, connections, who-walks-to-whom and the relative importance of nodes within a network (Barbier & Liu 2011; Fan & Gordon 2014). Chen et al. (2012) further discuss that recent research focusses especially on link mining and community detection. Link mining describes the process of discovering or predicting links between the nodes in a network while community detection or also called group detection is more interested in finding and identifying certain groups within a network. MentionMap is a social network analysis tool that is used to analyze conversations on Twitter and depict the tweets between Twitter users and their mutual relationships (Bloem et al. 2013) However, when analyzing data from social networks an important duty is to protect personal privacy especially since it has been found out that advanced data analysis techniques still may disclose personal information even though the information has been anonymized. Another downside of the topic of personal privacy is the problem that a person's privacy settings can interfere with the analysis and limit the ability of providing useful information (Barbier & Liu, 2011).

The next technique that will be introduced is visual analytics. Thomas and Cook (2006) define visual analytics as "the science of analytical reasoning facilitated by interactive visual interfaces" (p.10) whereas Keim, Kohlhammer, Ellis and Mansmann (2010) suggest the following definition based on the current application of visual analytics: "Visual analytics combines automated analysis techniques with interactive visualization for an effective understanding, reasoning and decision making on the basis of very large and complex data sets" (p. 7). Keim et al. (2010) further claim that the main goal of visual analytics is to make data processing and the resulting information as transparent as possible for an analytic discourse. Visual analytics is furthermore highly interdisciplinary and unites a number of different practices such as visualization, data mining, data management, data fusion, statistics and cognition science (Keim et al., 2010). Essentially it combines the strengths of data analytics with the human abilities of being able to quickly and visually realize patterns. Oelke et al. (2009) have used visual sentiment analysis on customer feedback and have introduced three main visual analytics techniques, in particular visual summary reports to provide a quick overview, visualization of clusters to present similar opinion groupings and circular correlation maps to detect correlations.

The practice of topic modelling uses a variety of advanced statistics and machine-learning techniques to study large bodies of collected textual documents to find an underlying dominant topic or theme which can give insights about customers' interests and emerging topics of discussion (Fan & Gordon, 2014). Especially Twitter data is often used in the area of topic modelling (Mathioudakis and Koudas, 2010; Hong and Davison, 2010; Zhao et al, 2011; & O'Connor, Krieger and Ahn, 2011). Mathioudakis and Koudas (2010) for example present the TwitterMonitor in their paper, which is a system that detects trends on Twitter by identifying trending or emerging keywords that are mentioned at an unusually high rate and then classifies them under a certain trend or topic. Based on these preceding steps, TwitterMonitor then uses tweets that are grouped under a certain trend and analyzes them further to filter out additional relevant information. Trend analysis is another technique in the field of social media analytics and describes according to Fan & Gordon (2014) a method that collects historical data and based on that data makes predictions about future outcomes and behavior such as forecasting sales growth. Social Media Analytics can yield a large potential of being a novel yet remarkable marketing tool in that it can provide marketers with information about consumers and finding patterns in customers' behavior in order to generate knowledge about what matters to them and what they could potentially need in the future. This opens up a whole new possibility for practitioners.

2.3 Organizational skills as an underlying strategic asset

In conjunction with the move to Social Media Analytics comes an increasing shift of awareness concerning the organizational skills that build the groundwork for efficient and successful Data processing efforts. With the intention of providing a clear outline as to how marketers and practitioners can profit from Social Media based forecasting, one must also address that a desire to gain insights about customers' needs or future market developments must be supported by the appropriate organizational requirements. This section therefore shortly discusses some of the organizational prerequisites. As an example, Fania and Miller (2012) point out that especially in the science of Big Data the skills that align the data with the business and that turn findings into positive business outcomes are essential. Additionally, Bloem et al. (2012) emphasize that Big Data can work as a stimulus to elevate a firm's business intelligence efforts to a drastically higher level with the appropriate technology, the right processes and the relevant skills and expertise to make efficient use of today's Social Media analytics possibilities.

One of the main conditions that has to be satisfied is the need for a clear strategy. In the era of Big Data almost every detail of data can be analyzed but in order to generate significant output from the analysis it is important to align corporate big data analytics and the overall corporate strategic goal. Fania and Miller (2012) lay out that the development of the organizational skills to mine and process Big Data to successfully execute predictive analytics can result in better decisions, increased business velocity, accelerates the pace of innovation and can aid in discovering and tapping new markets. In a survey of the Economist Intelligence Unit (2011) firms were asked about the two biggest challenges in extracting value from data and the second most common answer (30% of respondents) was a lack of the right skills within the organization to manage data effectively which illustrates again the importance of the right set of skills. In an effort to provide companies with an evaluation of their current position the Economist Intelligence Unit (2011) has created four loosely defined categories according to the level of Big Data management applied in the firm. These categories are as described below:

- *Data wasters*: These companies mostly collect data do not use it to their advantage. Additionally, a mid-level manager is commonly responsible for the firm's data strategy and there is often lack of alignment between business and information technology (IT).
- *Data collectors*: Companies in this category collect a large amount of data but do not manage them in a successful manner. There is no formal process or data governance and most likely a senior IT-executive is in charge of the firm's data strategy
- *Aspiring data managers*: These companies are aware of the value that can be derived from datasets. They are utilizing resources to take better advantage of the data they collected and base part of their decision making on information from data management However data strategy is not part of the CEO's range of tasks.
- *Strategic data managers*: This is the most advanced group within this categorization. Companies within this category have the most mature skills combined with well-defined data management strategies that are aligned with the data analysis and evaluation. A top level executive is in charge of the companies' data management strategy.

Firms can use this guideline to evaluate their current position and identify what changes might have to be implemented in order to ascend to the next higher level. This might mean that a company has to invest in the needed hardware, people and skills and eventually create clear management and governance strategies (Bloem et al., 2012). A variety of researchers have also highlighted the importance of so-called Data Scientists that go hand in hand with the increased complexity of Social Media Data (Harris & Mehrotra, 2014; Waller & Fawcett, 2013; McAfee & Brynjolfsson, 2012). Davenport and Patil (2012) define a data scientist as "a high ranking professional with the training and curiosity to make discoveries in the world of big data" (p.72). Data scientist differ in so far from traditional analysts as that in addition to numeric data they can also analyze unstructured, non-numeric data such as images, sound and text, they use different tools for example machine learning and open source tools such as Hadoop. These Data Scientists are therefore able to handle not only the enormous data deluge but also deal with the intricacy of Social Media Data in such a manner that they are able to sift out the answers to business problems and can give advice to executives and product managers concerning the implications of the data for products, processes and business decisions (Davenport & Patil, 2012). As a concluding remark it can be said that companies that fail to develop a Big Data competency are going to lag behind in the long run (Economist Intelligence Unit, 2011).

2.4 Limitations of Social Media based Forecasting

Even though many researchers are quite optimistic about the outcomes of their studies and sense positive appraisal about the predictive power of Social Media, there are also critics that are not as convinced for various reasons. In order to provide a comprehensive and critical evaluation of Social Media based forecasting and to avoid unintended consequences or false expectations this section of the literature will discuss the pitfalls and limitations of using Social Media Data as a forecasting tool. First of all it has to be mentioned that Metaxas, Mustafaraj and Gayo-Avello (2011) some of the main critics of social media based predictions, explain that most of the work that has been published especially concerning the prediction of election outcomes did not propose a method that would consistently be able to predict outcomes before they actually happen. They further assess the accuracy of lexicon-based sentiment analysis as quite poor when applied to political conversations as it performs just slightly better that a random classifier in that is cannot detect and correctly assign the intent behind disinformation and propaganda. Another skeptic of predictive analytics is Gary King, a Harvard University professor. His main concern is that people are constantly influenced by their surroundings, which are in fact unpredictable and cannot be known and measured accurately as a whole as they change continuously. This circumstance eventually signifies that "statistical prediction is only valid in sterile laboratory conditions" (Strickland, 2015; p.24). Lazer, Kennedy, King and Vespignani (2014) determine an additional core challenge of big data analysis which is the fact that the data that is being analyzed is coming from sources that are not the output of instruments designed to produce valid and reliable data that can be used for scientific analysis. Especially since major search engines are prominently showing trending topics, from real-time micro-blogging sites such as Twitter, in the search results people use certain tactics to promote their postings, a large amount of spam and political propaganda is being generated (Castillo, Mendoza & Poblete, 2012). Similarly, Schoen et al. (2013) put forward that Social Media Data can create additional difficulties with regard to the quality and credibility of the collected data which is to a certain extent attributable to users that use fake accounts, spread false information or use automated accounts that constantly update and thus creating an abnormally high volume of conversation. Noisy, false and incomplete data can distort the data analysis in such a way that the outcomes do not portray a veritable pattern. To overcome this kind of mistake research point to the complete process of knowledge discovery from databases in which data is cleaned out from outliers before it is thoroughly analyzed (Han et al. 2012.) Despite the fact that millions of users are active on social media, creating massive amounts of data every day, social media users still are not a representative sample of the population (Kalampokis, Tambouris and Tarabanis 2013 & Schoen et al. 2013). One explanation for this is the fact that the younger generation which is far more prevalent on Social Media is often over-represented in samples and therefore the results cannot signify the whole population (Gayo-Avello, 2011). Gayo-Avello (2012) adds that in general a demographic bias exists since not every age, gender, social or racial group is represented equally on Twitter. This can be generalized for most of the other Social Media platforms that are currently used. Another explanation could perhaps be that most users simply do not actively participate very much. Nielsen (2006) explains the participation inequality in Social Media on the basis of the 90-9-1 rule. This rule basically states that 90 percent of social media users are only reading and observing but will most likely not contribute anything. Nine percent of online users will contribute from time to time, resulting in ten percent of overall postings on the internet and lastly, 1 percent of social media users are responsible for 90 percent of all postings on Social media platforms. Gayo-Avello (2012) refers to this phenomenon as the 'silent majority' or 'silent observers'. Singh (1990) has furthermore, classified unsatisfied customers into four categories, 'passives', 'voicers', 'irates' and 'activists' which explains how different users will react very differently to dissatisfaction and shows that not everyone will share their opinion on the Internet. Moreover, a sampling bias in the data can skew results, which is especially true if satisfied customers remain silent while those customers with a more extreme position loudly express their opinion (Fan & Gordon, 2014). Lazer et al. (2014) further criticize that some researchers view Big Data as a substitute for traditional ways of data collection and analysis and ignore established standards of measurement, construct validity and reliability and dependencies among data instead of treating big data as a supplement to small data. Another important aspect to be considered is the fact that positive correlations and outcomes may be overly euphorically interpreted by researchers. Gayo-Avello (2011) calls this the 'file drawer effect' and warns about the drastic risks of ignoring negative results. The last criticism of using Social Media data is the area of personal privacy. Since the emergence of the Internet and especially Social Media, all kinds of personal information and data from everyday online activities are saved and stored which brings along the fear of loss of privacy, an anxiety among people that far more private information gets public than they would like (Bloem et al., 2013). Additionally, with an increasing amount of data analytics techniques and tools on the market there is an increasing concern that Data analysis may pose a threat to peoples' privacy and data security (Han et al. 2012). Especially when dealing with sensitive customer data it is therefore extremely important that firms have mastered the privacy regulations within the company to comply with the general privacy laws.

3. CONCLUSIONS

This section will discuss the findings of the literature review in order to evaluate the potential of Social Media data as a predictive tool and to be able to deduce a framework as to how the data can be used to forecast future needs and market trends. The first part of the literature review presented the main areas of interest in the area of predicting with social media data. Those areas include in particular the health care sector, politics, economics, social sciences and the stock market. In the health care sector Social Media data can reveal disease outbreaks and thus predict the further spreading of diseases. For politicians and governments Social Media can open up the possibilities to assess the public opinion and to evaluate and prognosticate election outcomes. In a similar manner, researchers can assess the development of the stock market by analyzing public mood and look out for emotional outbursts. Even though these areas do not specifically give insights as to how Social Media data can forecast market trends or customer needs, they provide a basis and indicate some of the pitfalls of predicting based on customer generated data. One major aspect that has to be highlighted is that the literature review revealed that earlier research in the particular field of elections did mainly not produce accurate predictions with regard to the fact that some of them could not replicated with different data sets and they did not actually predict any outcomes as "post processing of social media data has resulted in claims that they might had been able to make correct electoral predictions" (Metaxas et al., 2011; p. 166).

The largest sub-section of the literature review investigates methods and techniques that can potentially be used to analyze customer generated Social Media data to gain insights about customer needs and future market trends. A variety of techniques could be used but due to a time-constraint, only a limited amount of techniques were explored within this paper. The literature revealed two main fields of data processing which are Data Mining and Social Media Analytics. Data Mining consists of using statistical approaches and machine learning approaches such as cluster analysis and classification analysis to detect underlying or hidden patterns. Although these techniques can provide interesting information they might not be fitting enough to work with the complex and diverse data that is generated by customers on Social Media. Therefore a new field of research evolved which can be classified as Social Media Analytics. Within this topic five different techniques are presented, namely sentiment analysis, social network analysis, trend analysis, visual analytics and topic modelling. Since a sentiment analysis is a tool that has been used within various empirical studies it is also one technique that has received a large amount of critique. In order to function as precise and substantial as possible it is therefore worthwhile to propose some delicate but very beneficial adjustments. Based on the assertion of Gayo-Avello (2012) it seems to be a central task to improve the text analysis techniques to be able to detect humor and sarcasm in text documents. Furthermore is it necessary to execute a sort of credibility check so that propaganda and disinformation can be detected and excluded from the analysis. Researchers have further highlighted the importance of incorporating demographics and consider user participation and self-selection bias when performing a sentiment analysis.

The third section of the literature review aims at answering the sub-question which is concerned with the organizational skills that are needed to make predictions based on Social Media data. The literature has shown that there is a growing interest in this area as it is strategically important to have the right set of skills within an organization. It has been pointed out that Big Data and especially Social Media Big Data position firms in a situation where basic know-how is not enough to deal with the complexity of the data and the richness of differing analysis techniques which creates the requirement for firms to align the overall strategic goal of the firm with the appropriate organizational qualifications. To be able to create the most substantial outcomes it is advised to undertake and manage changes within the firm that include decision making, The literature moreover revealed a growing interest in recruiting so-called Data Scientists that are specialized in managing and analyzing large and complex data sets while making use of a variety of methods. The problem is however that Data Scientists are still a fairly new subject of interest and there is only a limited amount of actual Data Scientists readily available (McAfee & Brynjolfsson, 2012). This makes it hard for middle-sized and especially smaller firms to manage the data deluge and the complex nature of the data to extract useful and significant information from it in order to be able to predict market trends and customer needs.

The last sub-question is concerned with the criticism towards the predictive quality of Social Media Data and the limitations of Social Media based forecasting. The main limitations of Social Media based forecasting is the bias-nature of Social Media data and the non-representativeness of Social Media Data. Social Media Data is mainly criticized for being noisy, unclear and originating from platforms on the World Wide Web that were not build to produce output that is used for scientific analysis. It is therefore important to clean out spam, false information and propaganda to only analyses the relevant data. Furthermore, it is discussed that Social Media users are just a sample from the overall internet users and thus cannot be representative for an entire population. Moreover, a demographic bias exists in so far as age, gender, nationality, race or religion is not equally represented. Next to that, there is an asymmetry among users that actively participate and those that are regarded as the silent majority which is the large amount of users that simply observe. Ignoring the aforementioned limitations results in findings that are false and that provide an incorrect picture. Firms that base their decision making on the outcomes of a faulty predictive analysis might face drastic outcomes. Lastly, a large problem in the handling of customer data is the issue of privacy it is therefore of great importance to consider the need for security measures that prevent the disclosure of sensitive customer data.

After discussing the sub-questions it is now essential and possible to consider the main research question: *What is the value and predictive potential of customer generated Social Media data for firms in order to create business intelligence concerning customer needs, innovation and market trends?* From the literature review it can be derived that customer data from Social Media can in fact be used to make assertions about customer needs and market developments. Different approaches and techniques can produce different information outputs. A sentiment analysis can for example provide an insight about how a customer feels about a certain product and where the customer sees a potential for improvement by analyzing customer posts and conversations. Based on this information firms can make predictions about what needs are eventually grow to become a market need and can start to work on finding ways to close this gap. Trend analysis is another method that can help in identifying changing customer tastes and interests. Topic modelling can additionally be used to figure out what people are currently talking about by analyzing for example trending topics on Twitter, customer posts, conversations and hashtags. Information from trends or reoccurring topics can be used to understand what kind of innovations are needed in the market and thus to stay ahead of competitors. A social network analysis can help firms in segmenting customers into groups and finding out what specific groups for example a target demographic wants from a certain product or if they have differing needs. Social Media analytics can furthermore be used by firms to track the performance of competitors and by additionally analyzing what customers think about competitors' products firms can gain supplementary information about desired product features. If a firm is interested in making predictions about innovations it possible to concentrate on predicting market needs that will eventually lead to new innovations. Organizations can also modify existing tools to fit the purpose of interest. FoodMood.in can for instance be altered to show overall sentiment within one region about a specific company, product or service. This way firms could get an understanding about brand perception or how a certain region assesses a specific product or service either of the own brand or of competing firms. Again, based on this information organizations can unveil market needs and market trends.

Social Media analytics creates the opportunity to understand customers from a completely different perspective. The conclusion above shows that Social Media yields the possibility to provide critical business insights about customers but also about competitors. Nonetheless the discussed limitations still pose pending challenges that need to be overcome to make predictions more accurate and precise as possible.

3.1 Data Analysis Toolbox

There are a variety of tools that are being developed to help firms in making the most of the amassed customer data from Social Media activity. Not all existing tools can be discussed within the extent of this paper but a few of them will be shortly introduced to provide firms with an overview of the possibilities. Throughout section 2.2.1.2 some tools were already presented that made use of the particular technique that was explained. The following list will present tools that are used for the data collection, data management and data analysis of Social Media data.

3.1.1 Data Collection Tools

- API's
- FastField mobile forms

3.1.2 Database Management Tools

- HP Vertica
- OpenRefine

3.1.3 Data Analysis Tools

- HP Autonomy – Intelligent Data Operating Layer (IDOL)
- HP Distributed R
- Apache Hadoop
- Coosto
- SAS Enterprise Miner
- SAS Rapid Predictive Modeler
- IBM Business Analytics

3.2 Derived framework

In order to position this paper as a guideline for marketing strategy a framework has been derived to provide an overview and a guideline in how firms can analyze customer-generated Social Media data in order to make predictions about their customers' needs, market developments and market trends. The framework presented in this paper is based on the social media data analysis framework of Kalampokis at al. 2013 and on the Knowledge Discovery in Databases process. A complete overview of the Social Media Analytics process is presented in Figure 1. The generated framework consists of three phases which are named as follows: *Organizational background*, *Preparation* and *Data Analysis.* The process is iterative and the individual steps within the process can be repeated if it becomes apparent that for instance a mistake occurred or that the wrong measurements have been selected.

As the literature review has presented the appropriate organizational background as a requirement for successful Social Media Analytics implementation it is a basic prerequisite to have the suitable assets in the form of human resources, technology and management and regulatory frameworks. The first phase of the process is therefore concerned with the *Organizational background*. In order to receive a broad overview of the currents situation it can be advised for firms to position themselves in one of the four categories that were defined by the Economist Intelligence Unit (2011). Based on the placement, firms become aware of what changes might be necessary to efficiently exploit Social Media data and can arrange for the implementation of the required modifications. These moderations can include for example the acquisition of new technology to either store or analyze the data, recruiting specialized human resources that are experienced in handling complex Social Media data, establishing a new management framework and introducing new regulations for instance privacy regulations to prevent personal data to become public. A new management framework is in so far important and relevant as the transition to analyzing Social Media Big Data entails five management challenges that are leadership, talent management, technology, decision making and company culture (McAfee and Brynjolfsson, 2012).

After the organizational affairs have been put into effect the second phase, namely *Preparation* can be approached. The first step within the second part of the process is to decide on what it is that the firm would like to know i.e. the goal of the process. When this has been communicated and decided on, this step also involves the selection of the appropriate means of measurement which is an essential and very important step that can also positively influence the final outcome of the process. The next step is data conditioning which means to select target data from the large amount of raw data based on the desired outcomes. Firms could for example focus on Twitter data only or only use Social Media data from a certain region. Included in this step is also a pre-processing and cleaning stage which is implemented to ensure that noisy and irrelevant data is not included in the main analytical process. Cleaning out the data is highly important to create the most significant results as possible.

The last phase is the *Data Analysis* phase which also contains two steps. The first step is the actual Social Media data analysis. Following the preprocessing steps it is now important to decide for an analysis method. The literature review has presented Data mining techniques as the groundwork for Social Media analytics. From the list of data analytics approaches firms can use one or more methods to make sense of the complex data mountain and to unveil underlying patterns about their customers. Alternatively organizations can choose from the extensive analytics toolbox from which some tools have been introduced in the paragraph 3.1. Visualization methods can be used to further portray and to clarify the results of the analysis. The last step within the Data Analysis phase is concerned with the interpretation and evaluation of the analytical performance. It is evaluated if the selected measures have created significant and relevant outcomes. If the analytics process generated the desired information outputs the next step is to make sense of the outcomes and using the newly discovered insights within the firm.

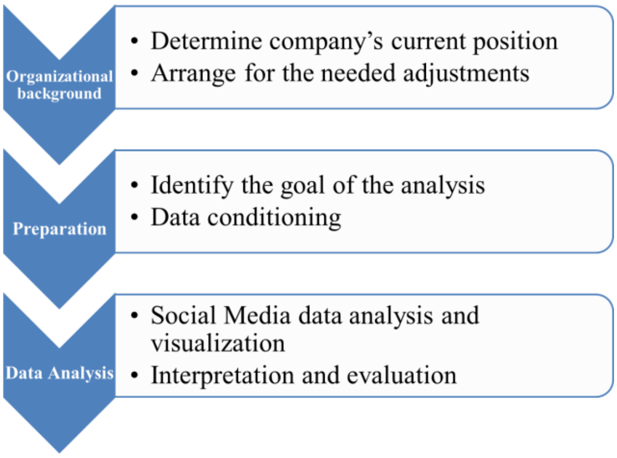
Organizational background
Determine company's current position
Arrange for the needed adjustments
Preparation
Identify the goal of the analysis
Data conditioning
Data Analysis
Social Media data analysis and visualization
Interpretation and evaluation

Figure 1. The derived Social Media Analytics framework

4. DISCUSSION

4.1 Final Remarks

This paper has reviewed scientific literature to create an overview of the predictive potential of Social Media data and to present a guideline as to how firms and practitioners can create business value on the basis of forecasting with Social Media Big Data. A new framework has been created that tries to minimize faulty outcomes that were based on the limiting factors of Social Media data. The framework furthermore incorporates important organizational prerequisites to ensure the successful application of the Social Media Analytics process. The review additionally presented a toolbox of professional tools that can be used to aid organizations. In conclusion it can be said that Social Media indeed provides the opportunity to predict market trends and customer needs with the utilization of Social Media Analytics however one has to be careful to not be too over-enthusiastic as there are a variety of limitations to Social Media data. Moreover, due to the fast changing pace in the World Wide Web new platforms arise constantly which introduce new difficulties in analyzing customer data from Social Media. It is therefore inevitable to stay up to date and to be able to continuously generate insights from emerging Social Media platforms.

4.2 Relevance

The paper is of academic relevance as it combines the existing research and literature into one paper and deliberates on the existing academic work. This literature review will therefore provide a clear overview of the state-of-the-art techniques of processing large amounts of Social Media data in order to predict future developments. Furthermore, this paper is intended to indicate any gaps in the current field of literature. The practical relevance of this paper is that it provides a clear overview, and evaluation of techniques that are currently used for forecasting based on Social Media data. This overview will not only offer a guideline for a firm's marketing department on what technique to use but also it will help companies in deciding how to pursue data processing of social media data. This literature review discusses limitations and pitfalls to the current techniques which will essentially aid firms in the decision and implementation phases of using customer generated Data from Social Media as in order to successfully exploit the presented techniques and avoid any negative outcomes, firms need to be aware that there are also limitations to predicting with social media data. The proposed framework is of great practical relevance as it recommends a guideline of how to successfully pursue the Social Media Analytics process.

4.3 Limitations and Further Research

The main limitation is concerning the time aspect. Due to a time-constraint it was not possible to find, analyze and review all relevant literature in this field. Not all techniques and methods that currently exist in the area of Social Media data and its forecasting potential and also not all existing tools that can be used for data management and data analytics could be investigated and introduced within the extent of this paper. The time limitation can also be used to explain that it would not have been possible to conduct an empirical study within the short period of about 10 weeks. An extensive study analyzing each of the discussed Data mining and Social Media Analytics approaches would be needed to verify and validate the findings of this literature review. Therefore, in order to successfully exploit the predictive potential of social media and in order to generate reliable results, deeper experiments and research has to be conducted in the future.

5. ACKNOWLEDGMENTS

Herewith I would like to thank my supervisor Dr. E. Constantinides for the continued and reliable assistance in realizing this paper. Additionally I want to thank J. van Benthem for evaluating this paper as a second supervisor.

6. REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S., Liu, B. (2011). Predicting Flu Trends using Twitter Data. *IEEE Conference on Computer Communications Workshops, IEEE;* pp. 702-707

Asur, S. & Huberman, B.A. (2012) Predicting the Future With Social Media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.*

Barbier, G. & Liu, H. (2011). Data Mining in Social Media. *Chapter 12 from the book Social Network Data Analytis. Springer* pp. 327-352

Barton, D. & Court, D. (2012) Making Advanced Analytics work for you. *Harvard Business Review, Vol. 90, Iss. 10*; pp. 79-83

Bloem, J., van Doorn, M., Duivestein, S., van Manen, T. & van Ommeren, E. (2013). No More Secrets with Big Data Analytics. *VINT Research Report* pp. 1-210

Bollen, J., Mao, H., & Zeng, X.J., (2010). Twitter mood predicts the stock market. Retrieved from: http://arxiv.org/pdf/1010.3003v1.pdf

Boutet, A., Kim, H. & Yoneki, E. (2012). What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*

Boyd, D. & Crawford, K. (2012) Critical Questions for Big Data. *Information, Communication & Society Vol. 15, Iss. 5*; pp.662-679

Castillo, C., Mendoza, M. & Poblete, B. (2012). Predicting information credibility in time-sensitive social media. *Internet Research Vol. 23 Iss. 5*; pp. 560-588

Chen, H., Chiang, R.H.L., & Storey, V.C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly – Special Issue: Business Intelligence Vol. 36 Iss. 4*, pp.1165-1188

Cooke, M. & Buckley, N. (2008). Web 2.0, social networks and the future of market research. *International Journal of Market Research Vol. 50 Issue 2*, pp.267-292

Corley, C.D., Cook, D.J., Mikler, A.R. & Singh, K.P (2010). Text and Structural Data Mining of Influenza mentions in Web and Social Media. *International Journal of Environmental Research and Public Health, Vol. 7, Iss. 2*; pp. 596-615

Davenport, T.H. & Patil, D.J., (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review Vol. 90 No. 10* pp. 70-76

Earle, P.S., Bowden, D.C. & Guy, M. (2011). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geography*, 54, 6, pp. 708-715

Fan, G. & Gordon, M.D. (2014). The Power of Social Media Analytics. *Communicatins of the ACM Vol. 57 Iss. 8;* pp. 74-81

Fania, M. & Miller, J.D., (2012). Mining Big Data in the Enterprise for Better Business Intelligence. Intel IT Best Practices- Business Intelligence; pp. 1-7

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Vol. 17 No. 3*; pp. 37-53

Franch, F. (2013). (Wisdom of crowds)^2: UK election prediction with social media. *Journal of Information Technology & Politics Vol. 10 No. 1;* pp. 57-71

Gandomi, A. & Haider, M. (2014). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management 3;* pp. 137-144

Gayo-Avello, D. (2011). Don't Turn Social Media Into Another 'Literary Digest' Poll. *Communications of the ACM Vol. 54 Iss. 10*; pp.121-128

Gayo-Avello, D. (2012). I wanted to Predict Elections with Twitter and all I got was this Lousy Paper – A Balanced Survey on Election Prediction using Twitter Data. Retrieved from http://arxiv.org/pdf/1204.6441v1.pdf

Ghose, A. & Ipeirotis, P.G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *Transactions on Knowledge and Data Engineering, Vol. 23 No.10,* pp. 1498-1512

Gilbert, E. & Karahalios, K. (2009a) Predicting Tie Strength with Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Science,* pp. 211-220

Gilbert, E. & Karahalios, K (2009b). Widespread Worry and the Stock Market. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature 45;*pp. 1012-1014

Goel, S. & Goldstein, D.G. (2014). Predicting Individual Behavior with Social Networks. *Marketing Science 33 (1)* pp. 82-93

Gundecha, P. & Liu, H. (2012). Mining Social Media: A Brief Introduction *Tutorials in Operations Research INFORMS, Vol.1 Iss.*4; pp. 1-17

Han, J., Kamber, M. & Pei, J. (2012) Data Mining: Concepts and Techniques. *Morgan Kaufmann 3rd Edition;* pp: 1-703

Harris, J.G. & Mehrotra, V. (2014). Getting Value From Your Data Scientists. *MIT Sloan Management Review Vol.58 Iss. 1,* pp. 14-19

Hildreth, S., & Ament, L. (2011). Benchmarking Enterprise Social Media Investment: Best Practices. *Hypathia Research*

Hoffman, D.L., & Fodor, M. (2010). Can You Measure the ROI of Your Social Media Marketing? *MIT Sloan Management Review (Fall)* pp. 41-49

Hong, L. & Davison, B.D. (2010). Empirical Study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics;* pp. 80-88

Jahanbakhsh, K. & Moon, J. (2014). The Predictive Power of Social Media: On the Predictability of U.S. Presidential Elections using Twitter. Retrieved from: http://arxiv.org/pdf/1407.0622.pdf

Jungherr, A., Jürgens, P. & Schoen, H. (2011). Why the Pirate Party won the German Election of 2009 or The Trouble with Predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. “Predicting Elections with Twitter, What 140 Characters Reveal About Political Sentiment”. *Social Science Computer Review 30 (2),* pp.229-234

Jungherr, A. & Jürgens, P. (2013). Forecasting the pulse. *Internet Research, Vol. 23 Iss. 5* pp. 589-607

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research, Vol. 23 Iss. 5* pp.544-559

Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, F. (2010). Solving problems with visual analytics. *Procedia Computer Science Vol. 7*; pp.117-120

Kietzmann, J.H., Hermkens, K., McCarthy, I.P. & Silvestre, B.S. (2011). Social media? Get serious! Understanding the building blocks of social media. *Business Horizons (2011) 54,* pp.241-251

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science Magazine, Vol.343*; pp. 1203-1205

Mathioudakis, M. & Koudas, N. (2010). TwitterMonitor: trend detection over the twitter stream. *Proceeding of the 2010 ACM SIGMOD International Conference on Management of data,* pp.1155-1158

McAfee, A. & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review Vol. 90, Iss. 10*; pp. 59-68

Metaxas, P.T., Mustafaraj, E, & Gayo-Avello, D. (2011). How (not) to predict elections. *IEEE Third International Conference on Social Computing;* pp. 165-171

Mejova, Y. (2009). Sentiment Analysis: An Overview. Retrieved from: http://www.academia.edu/291678/Sentiment_Analysis_An_Overview

Miller, G.A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM Vol. 38, Iss. 11*

Mishne, G. & Glance, N. (2006) Predicting Movie Sales from Blogger Sentiment. *American Association for Artificial Intelligence 2006: Computational Approaches to Analyzing Weblogs,* pp.155-158

Nielsen, J. (2006). The 90-9-1 Rule for Participation Inequality in Social Media and Online Communities. Retrieved from http://www.nngroup.com/articles/participation-inequality/ on 31.05.2015

O’Connor, B., Krieger, M. & Ahn, D. (2010). TweetMotif: Explanatory Search and Topic Summarization for Twitter. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media;* pp. 384-385

Oelke, D., Hao, M., Rohrdantz, C., Keim, D. A., Dayal, U., Haug, L., & Janetzko, H. (2009). Visual Opinion Analysis of Customer Feedback Data. *IEEE Symposium on Visual Analytics Science and Technology*; pp. 187-194

Pan, B. & Crotts, J.C. (unknown). Theoretical Models of Social Media, Marketing Implications, and Future Research Directions

Pang, B. & Lee, L., (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval Vol. 2 Iss.*1 pp. 1-135

Polgreen, P.M., Chen, Y., Pennock, D.M., & Nelson, F.D. (2008). Using Internet Searches for Influenza surveillance. *Oxford Journal of Clinical Infectious Diseases Vol. 47, Iss. 11*; pp. 1443-1448

Russom, P. (2011). Big Data Analytics. *TDWI Best Practice Report*

Sakari, T., Okazaki, M., Matsuo,Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *19th International Conference on World Wide Web (WWW'10) ACM Press,* pp.851-860

Schoen, H., Gayo-Avello, D., Metaxas, P.T., Mustafaraj, E., Strohmaier, M. &Gloor, P. (2013). The Power of Predicting with Social Media. *Internet Research Vol.23 Issue 5,* pp.528-543

Shmueli, G. & Koppius, O.R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly Vol. 35 Iss. 3*; pp. 553-572

Signorini, A., Segre, A.M., and Polgreen, P.M. (2011). The use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE, Vol. 6 No. 5*

Singh, J. (1990). A Typology of Consumer Dissatisfaction Response Styles. *Journal of Retailing, Vol. 66 Iss. 1*; pp.57-99

Strickland, J. (2015). Predictive Analytics with R. *Lulu.com 1st Edition*

Subasic, P. & Huettner, A. (2001) Affect Analysis of Text Using Fuzzy Semantic Typing. *IEEE Transaction on Fuzzy Systems Vol.9 No.4* pp. 483-496

Thomas, J.J. & Cook, K.A., (2006). A Visual Analytics Agenda. *IEEE Computer Graphics and Applications* pp. 10-13

Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. (2010). Predicting Elections with Twitter, What 140 Characters Reveal About Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media;* pp.178-185

Waller, M.A. & Fawcett, S.E., (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics Vol. 34 Iss. 2*; pp. 77-84

Wiebe, J.M., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning subjective language. *Computational Linguistics Vol. 30*; pp.277-308

Xu, J. & Liu, J. (2014). Forecasting Popularity of Videos using Social Media

Yu, S. & Kak, S. (2012). A Survey of Prediction Using Social Media. CoRR abs/1203.1647. Retrieved from: http://arxiv.org/abs/1203.1647

Zhang, X., Fuehres, H. & Gloor, P.A. (2011) Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia Social and Behavioral Sciences 26;* pp.55-62

Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E. Yan, H., & Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. Advances in Information Retrieval Vol. 6611, pp. 338-349

Economist Intelligence Unit (2011) Big Data. Harnessing a game-changing asset; pp. 1-31

American Marketing Association (2012) How to Leverage Big Data to Monetize Customer Experiences; pp: 1-9