Social Media and Forecasting: What is the potential of Social Media as a forecasting tool?

Author: Melina Barakos
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

In pursuance of retaining a competitive advantage on the market, businesses continuously ought to be ahead of time, meaning that they have to produce innovative products which respond to customer needs on a regular basis. This can only be accomplished if organizations are able to detect future market trends and customer needs. Social Media generated data offers the insights which are required to make predictions of future market trends and customer needs. Marketers have to be aware of the complexity of Social Media generated data as it can present various obstacles. Diverse data processing methods need to be applied in order to turn raw data into something meaningful and useful.

The purpose of this paper is to review research findings and results on the role of Social Media as a forecasting tool; the study is conducted on the basis of a critical literature review in order to give a clear impression of the potential and the value of Social Media data for forecasting purposes. It was detected that Social Media does have the potential of predicting future market trends and customer needs. Marketers however have to be cautious due to numerous limitations of Social Media data and the data processing methods which have to be applied. Furthermore since the topic of forecasting market trends and customer needs using Social Media has not been addressed specifically so far, this paper establishes a new framework tailored for predictions regarding future market trends and customer needs. Furthermore, professional tools which support the process of prediction are identified.

Supervisors: Dr. E. Constantinides & Dr. R. Loohuis

Keywords

Social Media Data, Forecasting, Social Networking Services, Data Mining, Sentiment Analysis, Innovation, Market Trends, Customer Needs.

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

1. INTRODUCTION

1.1 Background Information

In today's digitalized world, data appears to be the currency of the twenty-first century. The main source of data is generated and gathered online, primarily from customer's activities on Social Media platforms in which individual users communicate, share opinions and network with one another (Schoen et al., 2013). Businesses make use of tremendous quantities of data that is collected from Social Networking Sites such as Facebook, Twitter, Youtube, Flickr, Instagram, various Blog platforms and many others. The data collected from such sites can be referred to as *Social Media Big Data* (Lazer et al., 2009). According to Gundecha and Liu (2012) "Social media gives users an easy-to-use way to communicate and network with each other on an unprecedented scale and at rates unseen in traditional media" (p.2). It is easier to actively participate in Social Media rather than traditional media due to the fact that firstly, anyone can contribute (Yu & Kak, 2012), and secondly because the process of being involved in traditional media channels requires substantial means of time and input and because "it [social media] has torn down the boundaries between authorship and readership" (Zeng et al., 2010, p.13). Big Data collected from Social Media platforms reveals unique and easily accessible insights about customers' interests, habits and desires, over a large geographic spread at a significantly low budget compared to traditional data collection methods. The advent of Big Data in Social Media totally changed the depth and opportunities of analysis we had up to now into something much more powerful (Tufekci, 2014). Asur and Huberman (2010) claim that the content that is produced on Social Media platforms is especially useful due to its "ease of use, speed and reach" (p. 1) and because "social media is fast, changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry" (p. 1). This gigantic range of topics and interests as well as the large geographic scope make it possible for organizations to analyze every type of industrial sector, country, gender, personal profile and any other attribute in which they wish to expand their knowledge and expertise. According to Alexa Ranking, Facebook and Twitter are amongst the top 10 most visited websites universally (Alexa Ranking, 2015). Twitter has 302 million active monthly users with 500 million Tweets sent per day (Twitter, 2015) and a total of 645,750 million registered users by March 25, 2015 (Statistic Brain, 2015). Facebook has 1,44 billion monthly active users as of March 31, 2015 (Facebook, 2015). These statistics demonstrate Facebook and Twitters' incredible growth and impact as well as the potential value and large scope of the data collected from these sources.

Data generated from the above mentioned platforms provides excellent opportunities for marketers, economists and statisticians to predict market developments and customers' needs, established from Social Media data. Additionally, Social Media data has the potential of providing organizations with a strong marketing strategy with the ability of developing innovative products and services by meeting customers' wishes and needs. The power of prediction of Social Media has become a much-talked about topic in the previous couple of years with increasing popularity focusing on different aspects such as predicting elections, the stock market, diseases and many other variables which will be thoroughly discussed later on in this paper. A prediction mechanism from Social Media data could potentially be an enormously valuable tool to organizations if it is handled appropriately; it could deliver a competitive advantage as it can convey insights into efficiently meeting constantly changing consumer desires, communicated through Social Media (IBM, 2014). Ideally, if Social Media data is accurately used as a predictive tool, organizations may well produce and deliver their customers innovative products and services which fulfill their personal requirements and desires before they themselves have thought about them seriously.

1.2 Research Problem

In order to retain competitive advantage, businesses constantly need to be ahead of time, i.e. they are obliged to come up with brand new, innovative products regularly. In order to be capable of accomplishing this goal, it is necessary to predict market trends and customer needs with Social Media data. The reason why Social Media based forecasting as a marketing strategy concerning the prediction of future events and is still so rare is due to the complex process of turning raw Social Media data into something meaningful. The problem which firms are facing is that of inefficient evaluation of the data they have at hand, they essentially don't know how to make the most effective and practical use of the data in order to use it as a predictive instrument. Additionally, since this topic has not reached an advanced level of general knowledge, organizations may not know how beneficial Social Media generated data can be, what it can be used for and how it can be applied. Since this is a fairly untouched subject, one can't assure 100 per cent that the data from Social Media can always or even ever provide valid and valuable information about the future. Consequently, understanding the value and potential of prediction as a marketing strategy and the limitations which Social Media Big Data faces is vital when aiming to use the data as a forecasting device "in order to be successful and avoid false expectations, misinformation or unintended consequences" (Schoen et al., 2013, p. 528).

Thus the research goal of this paper is to accumulate the present findings and experiences of Social Media data as a forecasting device in the form of a critical literature review in contemplation of providing a clear overview of this issue. The author is attempting to extend the knowledge of the predictive power of Social Media data and to gain insights into discovering how valuable and how much potential the data offers for the (possible) prediction of future events, market trends, behaviors and customers desires and needs. Existing practices and techniques of data processing tools will be examined and critically reviewed so that one may find clear evidence about the value and quality as well as the best approach of effectively using the data collected for the above declared purpose. Hence, the following research question will be addressed: *What is the value and potential of social media generated data to organizations for the purpose of predicting future market trends and customer needs?*

Moreover, the following sub questions will be answered within the critical literature review:

Sub Question 1: *What is or has already been predicted with Social Media Data?*

Sub Question 2: *How can Social Media Data be analyzed and be transformed into meaningful Data? (Outline of forecasting models, tools, taxonomy of data processing methods)*

Sub Question 3: *Can innovation be detected through Social Media data?*

Sub Question 4: *What are the limitations of Social Media based forecasting according to critics?*

1.3 Relevance of the Topic

A paper in this explicit academic field is a valuable addition as there are limited numbers of academic articles regarding this matter which present an over-all guideline. This means that there are various articles which discuss the topic of Social Media and its predictive power; however most of them are about a specific industry, organization or variable. This paper attempts to provide a general recommendation for marketers concerning the practicality of Social Media generated data and it's potential in acting as a predictive device for future market trends and customer needs. The practical relevance of this paper is that it makes an effort to provide a suitable guideline for the marketing practices and strategies of an organization. It essentially demonstrates whether or not it is worth investing the effort (monetary and timely) of dealing with the complicated procedure of data processing methods. This paper attempts to demonstrate to organizations how to make valuable use of Social Media generated data and its predictive power to their advantage. Furthermore this general overview should assist in using Social Media data as a forecasting tool especially for market trends and future customer needs, since this topic has hardly ever been addressed in previous literature.

The paper is organized as follows. Firstly the methodology of this paper is outlined. Then, an extensive critical literature review will be conducted, in which key terms will be defined as well as answering the above outlined sub questions in order to contribute new insights about Social Media data and its predictive potential. After defining key terms, an outline of previous studies and experiments of predicting the future with Social Media generated data in different sectors will be given. Subsequently, the taxonomy of the methods and techniques used in data processing for forecasting will be analyzed and outlined as well as the limitations and critical views of Social Media based forecasting. Afterwards the question of Innovation through Social Media data is addressed. One of the last sections include an in depth discussion about the key findings which has the intention of answering the above stated research and sub questions. Afterwards a new framework is introduced with a new approach of using Social Media data for forecasting market trends and customer needs followed by an outline of professional tools which can support the process of Social Media based predictions. Lastly a conclusion will be given including the limitations of this paper along with suggestions for further research.

2. METHODOLOGY

This paper takes the form of a critical literature review based on the Emerald guide on 'How to write a Literature review'. The literature review systematically analyzes established, applicable findings and experiences from academic publications, conference proceedings and white papers in the field of Social Media and its potential to predict the future. This paper is purely based on literature. The main focus in this study lies on analyzing and outlining different taxonomies of data processing methods and the possible areas of prediction. This information will primarily be assembled from academic publications or conference papers which have a strong focus on Social Media, Social and Computer Science. This paper attempts to accumulate a range of relevant literature and opinions regarding the issue of Social Media data acting as a predictive tool, in order to reach clear and straight-forward conclusions, to fill the gap of knowledge and to provide a general overview concerning this field of knowledge. The criterion for the paper selection was the focus of referring to Social Media data (whether Twitter, Facebook or any other platform) as a predictive tool. The articles used in the literature review are derived from electronic search engines, primarily Google Scholar, ScienceDirect as well as the University of Twente online library. Furthermore, the reference lists of selected, relevant articles were scanned in order to access additional literature which was not found during the original search procedure. The most essential key search terms used with regard to discovering applicable literature were "Social Media (Data)", "predictions", "market trends", "forecasting", "data processing methods" and "innovation through Social Media". The papers were classified as relevant or as not relevant after glancing over the abstract and the research goal. Focus was also on the date the literature was publicized. Recent articles were favored; nevertheless most selected literature is not older than from 2009, since this topic is still relatively fresh and undeveloped.

This literature review is based on 53 academic articles/ research papers. Since the topic of Social Media based forecasting is fairly fresh, the majority of the articles that were analyzed are white papers or conference proceedings. Most articles stem from web based conferences such as the Conference of Computer Communications or the International Conference on World Wide Web. Furthermore specific Social Media and Data Mining Conferences were also part of the collection. Moreover, numerous articles stem from information systems, internet based or general business journals such as the Journal of Information Management, Internet Research and Business Horizon.

3. LITERATURE REVIEW

3.1 Definition of Key Terms

Key terms which will frequently be mentioned throughout this literature review, are defined in the following. The purpose of this small section is to ensure that the reader precisely knows what the author is referring to.

3.1.1 Social Media

As Social Media has been a significantly, widespread topic in previous years, plenty of definitions are available. A few of them will now be outlined. According to Kaplan and Haenlein (2010) "Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (p. 61). Kietzmann et al. (2011) argue that "Social Media employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss and modify user-generated content" (p. 241). Lastly, Constantinides & Fountain (2008) define Social Media as "a collection of open-source, interactive and user controlled online applications expanding the experiences, knowledge and market power of the users as participants in business and social processes" (p. 232). Although all of these definitions sound somewhat diverse, in essence they convey a similar message; Social Media consists of the following elements: openness, sharing, networking, communication, togetherness, co-creation or user-generated content. These characteristics are important for determining whether or not the future can be predicted with Social Media data.

3.1.2 Social Networking Service (SNS)

When speaking about Social Media "Social Networking Services" continually appears to be an accompanying term as it can be seen as a subcategory of Social Media. According to Yu & Kak (2012) a Social Network can be defined as a "Social structure comprising of persons or organizations which usually are represented as nodes, together with social relations, which correspond to the links among nodes" (p. 1). Social Networking Services (SNSs) are websites where users can register and form

their own unique profile. There, they can interact, communicate and network with other users, share experiences and photos, find users who have similar interests (and add them as a 'friend' to their own personal network), as well as forming discussion groups which leads to the well-known user-generated content (Ahn et al., 2007; Yu & Kak, 2012). The most popular Social Networking Sites in the United States in March 2015 (based on market share of visits) were Facebook as number one, followed by Youtube, Google Plus and then Twitter (Statista, 2015). This paper will mainly focus on Facebook and Twitter as these Social Networking Sites are most applicable for the purpose of this research due to the valuable and 'chatty' data which can be gathered from these communication channels. Table 1 presents various Social Media Platforms alias Social Networking Services, split into a number of different categories in order to display the large scope and data potential.

Table 1. Different types of Social Media Platforms alias Social Networking Services (Barbier&Liu, 2011, Gundecha&Liu, 2012, Gandomi & Haider, 2015)

Types	**Example**
Social Networks	Facebook, LinkedIn, MySpace, Googleplus
Blogs	Blogger, WordPress
Microblogs	Twitter, Tumblr
Social News	Digg, Reddit
Social Bookmarking	Delicious, StumbleUpon
Media sharing	Instagram, Youtube
Wikis	Wikipedia
Review Sites	Yelp, Tripadvisor

3.1.3 Data Mining

Data mining is predominantly important in predicting the future with Social Media Data. This is because Data mining techniques can bring a lot of precious insights about human conduct and communication (Barbier & Liu, 2011). Data Mining essentially consists of applying mining techniques to discover configurations or relationships which otherwise would have not been found. Barbier & Liu (2011) came up with a respectable and simple definition: "data mining is identifying novel and actionable patterns in data" (p. 328). Data mining techniques can help overcome typical problems with Social Media data which are for instance the size of the data set, the noisiness and its dynamic nature (Barbier & Liu, 2011). Gundecha & Liu (2012) came up with a similar definition: "…to effectively handle large-scale data, extract actionable patterns and gain insightful knowledge" (p. 1).

3.1.4 Forecasting

Since 'forecasting' is likewise an often declared term in this literature review, it is important to shortly provide the reader with a definition. The Business Dictionary defines forecasting as the following: "A planning tool that helps management in its attempts to cope with the uncertainty of the future, relying mainly on data from the past and present and analysis of trends" (Business Dictionary, n.d.). The words 'forecasting' and 'predicting' are both used interchangeably in this paper.

3.2 Social Media as predictive tool

Even though this field of interest is fairly fresh in the academic environment, there have been various experiments and published studies with a range of different outcomes which will briefly be drawn out in the following. It seems that forecasting with Social Media data is quite popular in different industries like the health sector, business or the movie industry. Asur and Huberman (2010) used Social Media (to be specific, Twitter) to forecast box-office revenues for movies, by observing the rates at which movie tweets are created. Similarly, Oghina et al. (2012) established a model to predict movie ratings using Social Media data by observing the quantity of likes and dislikes on YouTube in combination with written expressions from Twitter regarding a selected movie. Zhang et al. (2011) as well as Bollen et al. (2011) predicted Stock Market Indicators through Twitter by watching out for emotional outbreaks and general moods. Achrekar et al. (2011) claim to have predicted flu trends with Twitter data by making use of their own developed Social Network Enabled Flu Trends (SNEFT) framework which observes messages posted on Twitter with reference to flu indicators. Likewise Colutta (2010) found a model to forecast influenza epidemics, also by analyzing Twitter influenza-linked messages. Moreover, Goel and Goldstein (2014) attempted to predict individual behavior with Social Networks, did however find that there are also limits to the full prediction process. Another common prediction variable with Social Media data is Sentiment by undertaking a Sentiment Analysis (Nguyen et al., 2012; Bifet & Frank, 2010). The almost certainly greatest theme up to now regarding predictions using Social Media data is in the politics division, specifically in predicting election outcomes (Mejova et al., 2013; Metaxas et al., 2011; Sang & Bos, 2012; Boutet et al., 2012; Franch, 2013). Other predicting variables include sales forecasts (Liu et al. 2007) and crime forecasts (Wang et al. 2012 & Bendler et al. 2014). As can be seen, the models and zones of prediction have a wide range with numerous, diverse methods of analyzing and applying data. Unfortunately, no articles can be found which examine the use of Social Media data for market predictions and customer needs. Most of the above stated articles are assertive and optimistic regarding their findings and the use of Social Media data as a predicting tool, presenting it as a relatively simple and straight-forward procedure.

Nevertheless there are critics out there who are in fact questioning this simple representation of data analysis and usage. Daniel Gayo-Avello (2012) is one of those critics – he has published several articles concerning the above specified issue. He claims that no one has genuinely delivered a proper prediction so far – everybody argues they would have been able to accurately predict the correct results, fact is however, that all 'predictions' were publicized after the final real-world result had already been published. Furthermore he debates that the data collected on Social Media platforms could be biased which makes the results invalid or less valid (Gayo-Avello, 2012; Metaxas et al. 2011). This matter will be further considered later on in this literature review.

3.3 Analysis of Social Media Data: Taxonomy of Data Processing Methods

This is undoubtedly the most important fragment of the research problem this paper is attempting to solve and also generally concerning the topic of Social Media as a forecasting tool. Turning raw data into something meaningful and significant is essentially the most difficult part of the process. Organizations that wish to make Social Media based predictions should have a well-thought through Business Intelligence system which supports them at any point in time. Negash (2004) defines a Business Intelligence system as follows: "BI systems combine data gathering, data storage, and knowledge management with analytical tools to present complex internal and competitive information to planners and decision makers" (p. 178). This definition demonstrates that a Business Intelligence system essentially contains all the important elements which are

necessary for Social Media based predictions and thorough data analysis. The different data processing methods and ways of analysis which are outlined in the subsequent sections are crucial components of practical Business Intelligence systems. This section is split into a number of parts for the purpose of keeping a systematic overview. Firstly, the characteristics of Social Media data will briefly be investigated in contemplation of examining why the process of analysis seems to be such a complex task. Following on, various data processing methods will be outlined, closely inspected and evaluated with the intention of coming up with a guiding framework including an ordered means of using Social Media data for forecasting purposes. Subsequently, the question of Innovation from Social Media data is addressed. Lastly, critical views and limitations regarding the use of Social Media data for forecasting purposes are outlined.

3.3.1 Social Media Data Characteristics

Social Media delivers a low-priced, rapid and largely unstructured means of assembling data at a great scale and an extremely wide geographic scope (Schoen et al. 2013). Social Media data tends to appear mainly in the form of written contents (e.g. status updates on Facebook or Twitter, comments in reviews or social groups, conversations with other users etc.), but also in the form of likes or dislikes, tags, hashtags (particularly on Twitter and Instagram), emoticons, video messages, personal information (e.g. number of friends, citizenship, gender) and rating scores. Social Media data tends to be noisy, vast, distributed and informal in nature (Kalampokis et al. 2013; Barbier & Liu, 2011; Gundecha & Liu, 2012) since it originates from various mediums and in numerous different forms which suggests that it is predominantly very messy and unclear – consequently it is necessary that this raw data is transformed into qualitative, valuable data. Moreover, the dynamic nature as well as the absolutely enormous amount of data poses further challenges towards the use of Social Media data as a forecasting tool (Zeng et al. 2010, Barbier & Liu 2011). Examples of unstructured data comprise conversations, graphics, images, texts and videos – these can be turned into structured data by applying Data Mining techniques and other analytical processes (Negash, 2004). Essentially it is evident that Social Media data is difficult to analyze and evaluate due to four main characteristics: noisiness, size, unstructured and dynamic nature. These challenges need to be overcome in order to convert disordered and unclear data sets into something valuable to use them as a forecasting tool.

3.3.2 Taxonomy of Data Processing Methods

Scanning through existing literature, diverse methods of analysis can be encountered. Predictive Analyses are typically based on statistical approaches or data mining algorithms. Furthermore, Social Media analytics also play a great part; Zeng et al. (2010) define Social Media analytics as follows: "Social media analytics is concerned with developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application" (p. 14)

Schoen et al. (2013) introduce a taxonomy of models in their paper and point out three fundamental data collection and analysis practices, namely statistical models, survey models and (especially for Social Media data) prediction market models. With prediction market models, the behavior of users on Social Media platforms can be observed which can enable various predictions using the behavioral patterns that have been examined. Nevertheless, it has to be questioned, whether survey models are the correct way of collecting data from Social Media, due to a lack of accuracy and possible bias responses. Furthermore, the application of a Sentiment Analysis is also an often seen instrument in the data analysis process which investigates the thoughts and feelings of users, applying these to arrive at selected conclusions.

Gandomi & Haider (2015) suggest that the two most crucial information sources on Social Media platforms are User-generated content such as texts, videos and sentiments, and the relationships and interactions between the network participants. They claim that based on this classification of information sources, Social Media analytics can be split into two main groups, namely Content-based analytics and Structure-based analytics. Content-based analytics concentrate on the information which is published by users on Social Media platforms such as texts, comments, videos, images etc. Structure-based analytics on the other hand, focuses on the relationships between Social Media users (Gandomi & Haider, 2015). Similar to Structure-based analytics, Thiel et al. (2012) introduce a Network Analysis of Social Media data which also focuses on the relationships between individual users and their communication of various topics. Nodes and Edges are part of the representation of the structure of Social Media users and their relationships to one another. These relationships can be detected and examined through different types of graphs. Furthermore, Gandomi & Haider (2015) also outline various techniques that provide useful material from the structure of social networks such as community detection, social influence analysis and link prediction. These techniques are very useful for extracting Social Media based predictions. For instance, social influence analysis can investigate the sphere of influence a specific user has on a network, for example who acts as a leader and who acts as a follower in a specific network, implying that the leader has an increasing influence over the rest of the network (Thiel et al., 2012; Gandomi & Haider, 2015). Community detection can expose social patterns and lastly, links prediction techniques "predict the occurrence of interaction, collaboration, or influence among entities of a network in a specific time interval" (Gandomi & Haider, 2015 p. 143).

3.3.2.1 Statistical Methods

Gandomi & Haider (2015) suggest that "Predictive analytics techniques are primarily based on statistical methods" (p. 143). Constructing Statistical models to predict the future is always a useful technique to adopt when scrutinizing and extracting data sets (Truvé, 2011). Almost every piece of literature which was analyzed during this literature review, at some point made use of statistical methods in their empirical research or discussed statistical methods as a means of analysis for Social Media data in order to make sense of the data they have collected (Gilbert & Karahalios, 2009; Yu & Kak, 2012; Schoen et al., 2013; Tuarob & Tucker, 2013; Asur & Huberman, 2010; Lassen et al., 2014; Bandari et al., 2012; Achrekar et al., 2011). For instance, Jahanbakhsh & Moon (2014) analyzed political Tweets in order to predict the 2012 U.S. election outcome. In order to reach conclusions, they also applied statistical methods; they calculated the tweets frequency distribution, tweets mentions distribution and the hashtag distribution. The simplest and most frequent used technique is the Linear Regression Model which examines the relationship between the dependent variable, the prediction outcome and one or additional independent variables (Yu & Kak, 2012). These Statistical Models in essence prove whether or not there is a correlation or relationship between two or more variables. Statistical models act as enabling services which are part of almost every type of analysis, for instance Sentiment Analysis or Data Mining which will be discussed

deeper in the next sections. A benefit of applying statistical models for predictions is that one may choose out of a large range of predictors, whether they are 'tweets', Facebook comments and posts, number of likes or dislikes and many more (Schoen et al. 2013).

3.3.2.2 Data Mining with Social Media Data

Data Mining is an analytical procedure for the purpose of extracting and exploring large data sets which has the final objective to serve as predictive analytics (Statsoft, n.d.). The tremendous data sets from Social Media offer insights into Social Networks and the general society and Data Mining techniques make it possible to turn these data sets into something expressive and concluding. Data Mining has the potential of figuring out patterns, for instance common customer needs that are expressed on Social Media platforms which benefits organizations in terms of product developments as they can point out future trends early on in the process. Gundecha & Liu (2012) claim that Data Mining can "expand researchers' capability of understanding new phenomena due to the use of social media and improve business intelligence to provide better services and develop innovative opportunities" (p. 1) for example with techniques that "can help identify the influential people in the vast blogosphere, detect implicit or hidden groups in a social networking site, sense user sentiments for proactive planning, develop recommendation systems for tasks ranging from buying specific products to making new friends" (p. 1). According to Barbier & Liu (2011) data mining techniques are especially suitable in dealing with the main challenges that Social Media data bring along which have already been outlined above, namely its large, noisy and dynamic nature (Barbier & Liu 2011, Gundecha & Liu 2012). Typical areas for Data Mining in Social Media include community or group exposure, information diffusion, influence distribution, topic discovery and observation individual behavior analysis, group behavior analysis and market research (Barbier & Liu, 2011). In Data Mining, data is generally represented graphically, also with data from Social Networking Services – users are denoted as nodes and their relationships as links. Various Data Mining techniques and tools exist; however, it depends on the kind of data set which approach to make use of. The most common Data Mining applications related to Social Networking Services are outlined in the following. Barbier & Liu (2011) distinguish the types of approaches between Social Networking Sites and Blogs. The applications which are appropriate for Social Networking Sites according to the authors are group detection, group profiling and recommendation systems. The remaining methods which belong to Blogs are blog classification, identifying influential nodes, topic detection and change and finally Sentiment Analysis. Similarly, Gundecha & Liu (2012) also outline various Data Mining techniques, namely community analysis, Sentiment Analysis and Opinion Mining, social recommendation, influence modeling, information diffusion and provenance.

Group detection, group profiling and **community analysis** are similar to one another. Group detection is the application of spotting and recognizing a group within a Social Network by analyzing its structure and distinguishing individual users that are associated with one another. Marketers benefit from detecting the group an individual belongs to as they can gain insights into the users interests, habits and their social network. Hence, group profiling goes hand in hand with group detection as it is about classifying the profile of the group which was previously identified, i.e. what it is actually about. Barbier & Liu (2011) suggest that advanced Data Mining techniques can even detect changes in a group profile over a range of time "by defining a topic taxonomy" (p. 338). This can bring insights into how dynamic a group is and how rapid or how slow it changes its point of interest. Gundecha and Liu's (2012) Community Analysis is comparable with group detection and group profiling. The authors split the analysis into two parts; community detection and community evolution. Community detection is essentially the same as group detection; Gundecha and Liu (2012) define it as follows: "Community detection often refers to the extraction of implicit groups in a network" (p. 5). It is basically about the identification of various groups within a Social Network. Community evolution is similar to group profiling; however it is not 100 per cent the same application. The authors suggest that "community evolution aims to discover the patterns of a community over time with the presence of dynamic network interactions" (p. 5). This is similar to group profiling, especially concerning the detection of changes with topic taxonomy. It is more about distinguishing the changes of interests and direction of a group rather than identifying the characteristic of a group, which also belongs to group profiling.

Recommendation systems and **social recommendations** are both exactly the same application. In a Social Network context, such recommendation systems serve the purpose of increasing one's Social Network by suggesting for instance new friendships or group memberships to users who have mutual friends, interests or intentions, depending on their user profile (Barbier & Liu, 2011). This is carried out with automated data mining algorithms. Gundecha and Liu (2012) suggest that such recommendation systems are based on the premise that people who are connected with each other are expected to share similar interests or are easily influenced by one another.

Blog classification is a relatively simple application – it consists of the detection of different kinds of Blog profiles and types by applying data mining techniques. This means that blogs are automatically sorted for instance by topics (Barbier & Liu, 2011). The application of **Identifying influential nodes** involves the detection of influential bloggers or blogs which have the potential of gaining a large number of followers and supporters with certain products or messages (Barbier & Liu, 2011). This is especially valuable for marketers as this means that they can put their focus especially on such influential bloggers in order to detect market changes and customer interests as well as sentiments. Gundecha and Liu's (2012) **Influence Modeling** application is similar to Barbier and Liu's (2011) **Identifying influential nodes** application. Nevertheless, Gundecha and Liu (2012) distinguish between an influence driven and homophily (similarity) driven Social Network. The authors suggest that if a Social Network is homophily driven, a range of 'normal' individual users should be incentivized to promote products or services instead of influential bloggers. **Topic detection and change** implies that data mining techniques are applied in order to perceive topic trends and variations.

Sentiment Analysis and Opinion Mining are crucial applications in data mining. According to Gundecha and Liu (2012) "Sentiment analysis and opinion mining tools allow businesses to understand product sentiments, brand perception, new product perception, and reputation management" (p. 5). Essentially a Sentiment Analysis can perceive opinions and feelings which are expressed on Social Media platforms. Since this is a substantial topic of this paper, the next sub section will be solely dedicated to Sentiment Analysis and Opinion Mining.

Lastly **Information Diffusion and Provenance** is an analysis which examines how information spreads on Social Media platforms and which features affect the rate at which the

information disperses. Furthermore it attempts to discover the origin of the information; however this is a very challenging task due to the dynamic nature of Social Media data (Gundecha & Liu, 2012).

3.3.2.3 Sentiment Analysis & Opinion Mining

Making use of Sentiment Analyses by adopting data mining techniques is a common application in the literature which was examined. During the literature review it was discovered that such an analysis had been applied most frequently in the prediction of political election outcomes. Sentiment Analysis and Opinion Mining help to expose the emotions, feelings and opinions of users from various Social Media platforms by analyzing user-generated content such as text messages, posts, hashtags etc. Kim & Jeong (2015) refer to such content as consumer WOM (word-of-mouth). According to Chen and Zimbra (2010), Opinion Mining "refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated content" (p. 74). They suggest that Sentiment Analysis is used within the process of Opinion Mining in order to detect sentiments and feelings. This aids businesses with an improved understanding of their customers' feelings, needs and opinions in terms of their brand or products (Gundecha & Liu, 2012). Asur and Huberman (2010) define Sentiment Analysis as "a well-studied problem in linguistics and machine learning, with different classifiers and language models" (p. 497). To present an example, Liu et al. (2007) examined the predictive power of opinions and sentiments expressed in blogs in order to determine whether sales performance could be forecasted. They proposed a new model, namely 'Sentiment Probabilistic Latent Semantic Analysis' (S-PLSA). Instead of considering all words in a specific post or comment, S-PLSA concentrates only on sentiment-related words or phrases. Liu et al. (2007) claim that "since what the general public thinks of a product can no doubt influence how good it sells, understanding the opinions and sentiments expressed in the relevant blogs is of high importance, because these blogs can be a very good indicator of the product's future sales performance" (p. 607). In fact, this is a valuable point, not only for future sales performance, but likewise for any other predicting variable one may wish to foresee. This would also be an efficient way of predicting market trends and customer needs as it is essentially about what the customer feels and thinks about certain aspects and their desires. It was found that the content of sentiments expressed in entries or posts, is more influential than simply using the volume of blog (or from any other Social Media platform) mentions (Liu et al., 2007; Chung & Mustafaraj, 2011). Bermingham and Smeaton (2011) on the other hand argue that volume is a stronger indicator than sentiment due to the fact that volume is a practical indicator of popularity and because sentiment is ambiguous, making it difficult to "discriminate between sentiment which reflects the inner preferences of people, and that which is reflecting an immediate response to a given news story or event" (p. 9). Asur and Huberman (2010) also successfully made use of a Sentiment Analysis for the purpose of predicting box office revenues. They argue that when analyzing the sentiment of a tweet, box-office revenue predictions are enhanced. Tuarob and Tucker (2013) offer a Knowledge Discovery in Databases (KDD) model and in their study they attempted the prediction of product market adoption, also by analyzing sentiments of tweets. In order to categorize the type of sentiment they extracted from a tweet, the authors classify a tweet into three categories: positive, neutral and negative. This is the typical classification of sentiments which is used for analysis (Thiel et al., 2012). Similarly, O'Connor et al. (2010) used a Sentiment Analysis method in their study in which they assembled their day-to-day sentiment scores by counting positive and negative messages. They classified positive and negative words with the subjectivity lexicon from OpinionFinder. There are various linguistic analysis packages which are very useful for this type of analysis. Furthermore, Mejova et al. (2013) examined the sentiments which were expressed about politicians by developing a data-driven political sentiment classifier. Using the classifier, it enabled the authors to track changes in sentiments. Kim & Jeong (2015) claim that there is a lack of real-world social analytics for social media data; due to this, they established a practical approach. Their Social Opinion Mining Methodology consists of four stages: **1. Data Collection** (from API, crawling etc.), **2. Natural Language Processing** (removing unusable content, transforming data into something meaningful), **3. Data Analysis** (Sentiment and Statistical Analysis, Topic and Buzz Analysis etc.), **4. Presentation** (effective visualization of results with graphs, density heat map, valence tree map etc.). Sentiment on Social Media platforms is an important aspect to keep in mind as the thoughts and feelings expressed by users play a crucial role in their end decision whether or not to buy a product. It is the same game regarding forecasting purposes. The analysis of sentiments is a challenging task as the content which is analyzed i.e. language is very vague and can have numerous meanings or implications.

3.3.2.4 Social Influence Analysis

As aforementioned in section 3.3.3.2 (Data Mining), influence distribution is especially interesting for Social Media based predictions and has been discussed numerously as a predicting mechanism (Anagnostopoulos et al., 2008; Crandall et al. 2008). Social influence "refers to the phenomenon that the action of individuals can induce their friends to act in a similar way" (Anagnostopoulos et al., 2008 p. 8). This can be particularly applied to Social Media platforms such as Instagram. Instagram is a Social Media platform in which a user can post photos and videos with different captions and hashtags. These photos and videos can simultaneously be posted on various other Social Networking Services such as Facebook, Twitter and many others. Influential 'Instagrammers' with many followers, let's say approx. 100,000, often get noticed by brands which reach out to them for advertisement purposes. Taking fashion bloggers as an example; Chiara Ferragni is one of the most successful fashion bloggers of today with her blog called *The Blonde Salad*. Ferragni has about 3,9 million Instagram followers (Instagram: Chiara Ferragni, 2015) and sponsorship advertisement deals with various luxury fashion labels. This is an ideal example of Social Influence as many followers of Ferragni admire her style and wish to look like her by buying the same or similar products. Since most of those products are very expensive, this acts like a sort of predictive mechanism for cheaper brands to copy and develop high fashion trends for an affordable price. An analysis of Social Influence is a useful technique for the detection of future market trends and customer needs. Such bloggers or influential users which are active in Social Networking Services usually have products which are right in the beginning of their product lifecycle and well before they are commercialized.

3.4 Innovation through Social Media

Social Media based market trend predictions or product predictions have the potential of delivering innovative product developments. Unfortunately this is not a much discussed topic on the academic grounds so far, neither theoretically nor

empirically, but innovation through Social Media is a highly interesting field of study. Plessis (2007) defines Innovation as follows: "Innovation as the creation of new knowledge and ideas to facilitate new business outcomes, aimed at improving internal business processes and structures and to create market driven products and services" (p. 21). However, in the Social Media context we don't simply refer to this phenomenon as *innovation* but instead as *social innovation* (Charalabidis et al., 2014). Charalabidis et al. (2014) define social innovation as "a novel set of activities, performed by various social actors, such as government agencies of various layers (e.g., municipalities, regions, ministries), non-government organizations, firms, civil society, citizens' initiatives, or even individual citizens, entering in new forms and networks of cooperation, in order to address a problem not addressed by existing market offerings or government services" (p. 225). Social Innovation depends upon Social Networking i.e. user-generated content, communication, interaction and collaboration on Social Media platforms. Social Media acts as an excellent platform for Social Innovation as it enables the exchange of ideas and opinions amongst numerous different users and because it serves as a channel in which users can express their unfulfilled needs and desires. Charalabidis et al. (2014) suggest that when using various kinds of Social Media platforms, different types of users are attracted and therefore contribute to the innovation process which maximizes the input due to the range of different types of characters and opinions. In 2011 Kalypso, a U.S. innovation consulting firm, surveyed over 90 manufacturing and service companies to detect how Social Media impacts their product innovation process (Kalypso, 2011). **The survey found that over one half of questioned companies are making use of Social Media in product innovation to some point**. Those companies that are using Social Media for product innovation are "gaining business benefits, including more (and better) new product ideas or requirements, faster time to market, faster product adoption, lower product costs, and lower product development costs" (Kalypso, 2011, p. 3). Sine this is not a well-studied topic, businesses unfortunately are mostly overwhelmed which such a complicated task due to a lack of understanding. With the results of their survey, Kalypso (2011) identified various advantages of using Social Media for Product Innovation; *more* and *better* new product ideas or requirements, faster time to market, faster product adoption and lower product development costs. With the use of Social Media data, businesses are not only able to forecast future trends and customer needs, but also to detect innovations which essentially fulfill currently unsatisfied desires of their customers. Furthermore, businesses are not only capable of spotting innovations which they might not have come across, but they also save financial and timely resources with the use of Social Media data.

3.5 Criticism and Limitations of Social Media based forecasting

Since this paper is a critical literature review, the downsides of Social Media data have to be considered. While most of what has been said about Social Media based predictions in academic literature sounds promising and beneficial, the limitations and criticisms have to be pointed out so that caution is taken when using Social Media data for forecasting purposes. One researcher, namely Daniel Gayo-Avello is especially critical regarding the subject of Social Media based predictions, specializing on political election forecasts. His position regarding this matter is relatively simple: "*No, you cannot predict elections with Twitter*" (Gayo-Avello, 2012, p.2).

Social Media data is characterized as noisy and unclear, and the perhaps most discussed limitation of Social Media data is its **bias nature**. The users of Social Media are just a sample of all internet users i.e. they don't represent our entire population in general. Not everyone is using Social Media which is why the sample from which data sets come from is most likely a very biased one (Gayo-Avello, 2011; Jahanbakhsh & Moon, 2014). Likewise, Gayo-Avello (2011) suggests that researchers suffer under a so called 'file drawer effect' – this is an inclination towards reporting positive outcomes after a couple of confident results while ignoring or suppressing the negative ones, which is another form of biased behavior and can have drastic consequences. Gayo-Avello (2011) furthermore argues that Social Media users are frequently represented by the younger generation which is again a form of bias – it is necessary to have an equal spread of all ages in order to have a reliable and representative sample. For example regarding election outcomes, Gayo-Avello (2011) proposes that the younger generation usually points towards a liberal political opinion, which is why this political direction gives the impression of appearing stronger on Social Media platforms (assuming that the younger generation represents the majority of Social Media users). Furthermore, Wijnhoven & Bloemen (2014) address the issue of external validity in Sentiment Mining as they argue that it is an important issue to consider due to due various biases. They base their research on Shadish et al.'s (2002) five threats to external validity (p. 86-90) namely the properties of sample units (T1), differences in treatments (T2), findings of specific studies that can't be generalized (T3), observations that could be bias due to specific settings (T4) and lastly the way in which causal patterns are detected (T5). Wijnhoven & Bloemen (2014) pair each external validity threat with possible biases in Sentiment Mining; T1 is paired with demographic bias, T2 with manipulations of reviews, T3 with bias caused by events, T4 with platform bias and T5 with algorithm bias. Essentially, self-selection bias and demographics are merely disregarded which can have fatal consequences in the future (Schoen et al., 2013; Gayo-Avello, 2012). To solve the challenge of self-selection bias, Schoen et al. (2013) propose that for instance, demographic bias "could be reduced by weighting contents accordingly to the strata to which each user belongs by using user profiling" (p. 8). This undoubtedly sounds a lot simpler than it essentially is in reality.

Similarly, various researchers criticize the **non-representativeness** of social media users and its data due to a variety of reasons. Firstly, fake accounts flaw data sets, manipulate them and make them unrepresentative for analysis. Fake accounts or 'rioters' that are active on Social Media platforms have the potential of completely contaminating the data collected by diffusing propaganda, making it misleading and consequentially useless (Metaxas et al., 2011; Metaxas & Mustafaraj, 2012). Bloem et al. (2012) state in their research paper that almost 9 per cent of Facebook accounts are not real (Facebook has admitted this) which totals to 83 million accounts. This is a huge amount which can have a drastic, meaningful impact on the data. Moreover, Tufekci (2014) also considers the question of representativeness and methodological pitfalls of Social Media data. She also argues that there is a sort of bias and non-representativeness of Social Media data due to the fact that most researches so far have only considered one platform, mostly Twitter. She claims that Twitter "lacks some of the characteristics that blogs, LiveJournal communities, or Facebook possess, such as longer texts, lengthier reaction times, stronger integration of visuals with text, the mutual nature of "friending" and the evolution of conversations over longer periods of time" (p. 507). Tufekci (2014) contends that it is not

sufficient to simply gather data from one platform as "Information in human affairs flows through all available channels" (p. 507) and not only one, which is why many platforms have to be taken into account when aiming for a representative sample of data. Another contributor to non-representativeness are users that are called 'silent observers' or 'silent majority' (Gayo-Avello, 2012) who simply watch what is happening on Social Media platforms, but rarely or even ever contribute. The 90-9-1 rule addresses this problem; 90 per cent of Social Media users are referred to as 'lurkers', 9 per cent contribute occasionally and only one per cent of users are actively contributing on a regular basis (Nielsen, 2006).This principally shows that the data that is collected only constitutes of maximum 10 per cent of Social Media users' views and opinions. Likewise Tufekci (2014) also states that it is not sufficient to know how many people contributed to Social Media actions, without knowing how many people essentially saw the post (or any other type of action) and essentially decided not to contribute.

Additionally, Sentiment Analyses, that rely on polarity lexicons in which terms are either classified as positive or negative, are depicted as "naïve classifier" (Gayo-Avello, 2011, p. 10). Metaxas et al. (2011) claim that the results of their study regarding electoral predictions "show that the accuracy of the sentiment analysis is only 36.85%, slightly better than a classifier randomly assigning the same three labels (positive, negative, and neutral)" (p. 168). The authors suggest that when relying on polarity lexicons when making use of a Sentiment Analysis, propaganda and misinformation are not only overlooked but also incorrectly interpreted, i.e. wrongly classified into one of three categories, namely positive, negative or neutral. Metaxas et al. (2011) conclude that a Sentiment Analysis is "a far cry from being able to predict political preference" (p. 169) mainly due to incorrect interpretations and understandings of expressed sentiments. Since this is a universal problem, election predictions as well as any other sort of prediction will expectedly be affected by this.

4. DISCUSSION OF FINDINGS

In this section the findings of the above analyzed literature are evaluated in order to be able to arrive at conclusions whether Social Media data is a trustworthy source for predictions. The research question this paper attempts to answer is '*What is the value and potential of social media generated data to organizations for the purpose of predicting future market trends and customer needs?'* In order to systematically answer this question, the main research question was split into a number of sub questions. These are going to be answered in this section with the final objective of answering the overall research question.

The first sub question which has to be addressed is *What is or has already been predicted with Social Media Data?* This is essentially an introductory question so that the reader is able to get acquainted with the topic of Social Media based forecasting and to get a feeling of what has been studied in this academic field so far. It was found that various studies were conducted, empirically as well as theoretically, in order to find out more about the predictive power of Social Media data. It was discovered that most of the previous studies that had been conducted regarding Social Media based forecasting are concerning political electoral outcomes. However the views regarding the outcomes are diverse. Other predictor variables that were found include box office revenues for movies, movie ratings, stock market indicators, flu trends, individual behavior, crime and sales forecasts. As can be seen, there is a widespread variation of predictor variables. Unfortunately no previous work was found which addresses the research question of this paper i.e. regarding future market trends and customer needs. This is perhaps due to the fact that it is a fairly vague field of study which makes it quite difficult.

For this paper, the most important and most relevant sub question is *How can Social Media Data be analyzed and be transformed into meaningful Data*. The answer to this question consists of techniques and data processing methods which are outlined and evaluated, that serve the purpose of transforming data into something meaningful by revealing patterns and new discoveries such as common customer needs and interests. A diverse range of data processing methods and techniques were identified in the literature that was studied. The literature review outlines the basic approaches and applications of data processing methods, unfortunately, explicit techniques could not be investigated with great detail due to a lack of time. As was discovered, what makes Social Media data so challenging to analyze and to use for forecasting purposes is due to its characteristics. It was found that Social Media data sets are categorized as rapid, unstructured, vast, distributed, noisy, informal, dynamic and large (Schoen et al., 2013; Kalampokis et al., 2013, Barbier & Liu, 2011; Zeng et al. 2010; Gundecha & Liu, 2012). These characteristics all point into a negative direction, which makes the process of analysis rather tough.

It was noticed in the literature that most data processing methods rely on the application of Statistical Models. The use of Statistical Models emphasizes the results and essentially 'proves' that there is a correlation or a relationship between two or more variables. Most researchers agree that one of the most crucial applications for the analysis of Social Media data for forecasting purposes is Data Mining. Data Mining is a wide-ranging concept which consists of various techniques. The reason why Data Mining is so useful for Social Media based predictions is because it has the potential of figuring out patterns, for instance common customer needs that are expressed on Social Media platforms which can result in upcoming trends and innovative product developments as organizations are able to point out future trends early on in the process. The most appropriate applications that were detected in Data Mining related to Social Media forecasting are group detection and group profiling aka community analysis, (social) Recommendation systems, Blog classification, the identification of influential nodes (Influence Modeling), topic detection and change, Sentiment Analysis and Opinion Mining and Information Diffusion and Provenance (Barbier & Liu, 2011; Gundecha & Liu, 2012). These applications are crucial for forecasting means as they disclose core topics and changes which can reveal trend patterns, customers' opinions and feelings regarding trends or products, the revelation of influencers and trend-setters, the characteristics of a group and its profile and how rapid and how information spreads through Social Media channels. Sentiment Analysis one of many applications of Data Mining, however it is such a crucial and frequently used technique in the analyzed literature regarding Social Media based predictions, that a separate section was dedicated to this method. Nothing is more important for businesses, than taking into account the opinions and feelings of customers. It is such insights which essentially give organizations a competitive advantage as they can act as a guide for decision making due to the exposure of emerging trends in the form of common topics and interests, customer needs, emotions and opinions. A Sentiment Analysis is carried out with a subjectivity lexicon which classified words as positive, negative and neutral. Although this may sound like a flawless technique for analyzing customers' emotions and opinions, Sentiment Analyses unfortunately have its downsides;

Wijnhoven & Bloemen (2014) criticize its external validity, arguing that its outcomes are bias. Furthermore, the reliance on such lexicons is referred to as naïve (Gayo-Avello, 2011) and that its' results are not accurate (Metaxas et al., 2011). An additional analysis which has not been greatly discussed in a Social Media based forecasting context so far and can also be referred to as a Data Mining application as was pointed out earlier on, is a Social Influence Analysis. Due to the vast amount of Social Media channels it is vital to detect the most influential users which act as leaders and trend-setters. This is especially important in identifying and forecasting upcoming market trends and customer needs as it can lead researchers into the right direction for forecasting purposes.

The subsequent sub question addressed the issue of innovation and the development of novel products through Social Media data. It was perceived and concluded from previous research that Social Media data does indeed have the potential of delivering fresh ideas from user-generated content as it can take account of customer needs and of current gaps on the market. A U.S. consulting firm found a range of advantages in a survey they conducted of using Social Media for product innovation such as improved new product ideas or requirements, swifter time to market, quicker product acceptance and inferior product development costs (Kalypso, 2011). Ultimately, the simplest answer to this sub question is yes, Social Media data does indeed have to potential of delivering product innovation to organizations by taking into account customer generated contents which include the opinions and stimulation from customers for new or improved product ideas.

The last sub question focuses on the limitations of Social Media based forecasting and its' accompanying data processing methods. Unfortunately it is not as easy as it seems to make Social Media based predictions. Various limitations have to be considered and taken into account to circumvent false expectations and to prohibit incorrect beliefs. When attempting to forecast trends or any other variable with Social Media data, one has to bear in mind the following factors: it is vital that the bias nature and non-representativeness of Social Media data and users is always considered when making predictions (Gayo-Avello, 2011; Metaxas et al., 2011; Metaxas & Mustafaraj, 2012; Bloem et al., 2012). Anyone, anywhere on this entire planet can be active on those Social Media channels, can express their opinion, can create false accounts and has the potential of completely misleading and misinforming researchers by contaminating their data sets. Furthermore, as already pointed out above, Sentiment Analyses also bring various limitations with it, namely the reliance on polarity lexicons used for analysis as Social Media content is often falsely interpreted and wrongly classified (Gayo-Avello, 2011; Metaxas et al., 2011).

As all sub questions have been considered, the overall research problem can now be addressed. The research question of this paper is '*What is the value and potential of social media generated data to organizations for the purpose of predicting future market trends and customer needs?'*

Having thoroughly analyzed the literature in this field of knowledge the question can be answered as follows. Social Media generated data does indeed have the potential of making predictions, also regarding market trends and customer needs. Social Media data is very precious as it captures customer insights such as their needs, emotions, unfulfilled wishes and market gaps. These insights are retrieved from customer generated contents such as status posts, hashtags, conversations, group entries, friendships, images and videos, profile information and many more. This information is very valuable, especially for predicting future market trends and upcoming customer needs. However, the process of prediction needs to be carried out with caution; as was pointed out in the literature review as well as in this section, there are numerous limitations to Social Media generated data and its data processing methods which can be applied for analysis. These limitations need to be taken into account by marketers when using Social Media generated data for prediction purposes. When considering the limitations with caution, Social Media based predictions can benefit organizations enormously as it can prosper their product developments through innovative ideas from customer inputs from Social Media platforms. This in turn gives organizations a competitive advantage as they are listening and responding to their customers and their needs (which they may not even have known about themselves).

5. FRAMEWORK FOR PREDICTING FUTURE MARKET TRENDS AND CUSTOMER NEEDS WITH SOCIAL MEDIA DATA

During the literature review it was perceived, that so far no proper framework has been established for Social Media based predictions regarding future trends and customer needs. This may be due to the fact that in this specialized field of study, no thorough research has been conducted so far. Kalampokis et al. (2013) offer a useful framework called 'The Social Media Data Analysis Framework' which acts like a step by step guide for using Social Media Data as a forecasting tool. The framework consists of two main phases; firstly the *Data Conditioning Phase* and secondly the *Predictive Analysis Phase*. The *Data Conditioning Phase* consists of the following steps: 1. Collection and Filtering of Raw Data. 2. Computation of Predictor Variables. 3. Creation of Predictive Model. The *Predictive Analysis Phase* consists of the Evaluation of the Predictive Performance. Although their framework seems like a great guide at a first glance, it seems to miss out on the limitations which Social Media Data brings along, most importantly as already pointed out above, the bias nature and non-representativeness of the data. Furthermore, this framework seems to be very broad. This paper proposes a framework based on Kalampokis et al. (2013) 'The Social Media Data Analysis Framework' however customized to market trend and customer needs predictions, taking into account the limitations which were outlined in the literature review.

Figure 1 presents a framework, established for Social Media based predictions regarding market trends and customer needs. The framework is split into four stages which will now be outlined in greater detail.

Stage one deals with the identification of the predictor variable. This implies that the researcher needs to think about what he actually aims to predict. This can be range of factors starting with general trends for instance upcoming trends in the fashion industry, mobile or technology industry. Nevertheless, this

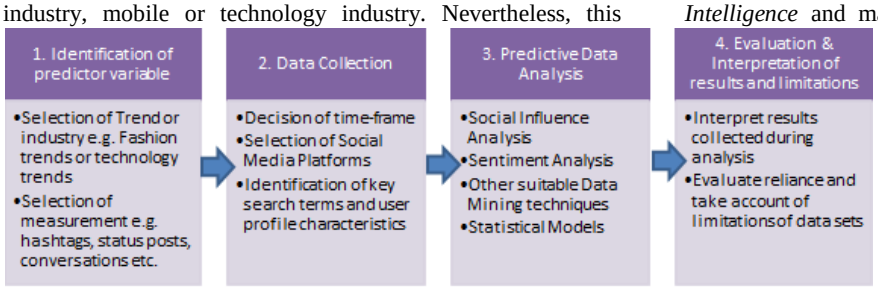

Figure 1: Framework for the prediction of market trends and customer needs with Social Media Data (based on Kalampokis et al., 2013)

predictor variable can be also be more detailed, for instance the marketer could choose a specific product and conduct research on customer needs on how to improve the quality of the product or to find out what factors have to be considered when developing innovative products which can be associated with existing products i.e. to fill the market gap which customers are currently facing. Additionally, at this stage the decision of measurement should also be approached. This simply means that the researcher has to decide which kinds of user-generated content he wishes to analyze. This can be a wide range of selected measurements such as hashtags, status posts, conversations, group entries, reviews, images, videos etc.

Stage two handles the process of Data Collection. During this stage the marketer has to decide where the data should be collected from i.e. one or more Social Media platforms have to be selected for the process of data collection. This can depend upon various factors such as age group, nationality, gender etc. This is also why the researcher may select specific user profile characteristics for analysis may filter out non-suitable data sets. Furthermore, the time frame of the collected data has to be specified as well as which key terms are selected for analysis.

Stage three manages the most complex part; the process of analysis. This framework proposes that a Social Influence Analysis, a Sentiment Analysis as well as Statistical Models should definitely be part of the procedure as these data processing methods are vital for detecting the required results as was pointed out during the literature review. Moreover, various other kinds of Data Mining techniques can be applied, depending on what it suitable for the predictor variable.

Lastly, **stage four** comprises the interpretation and evaluation of the results which were gathered during the process of analysis. This includes a critical assessment of the results, taking limitations, validity and reliability of the data processing methods as well as the data sets into account in order to avoid false predictions or expectations. The tough part about Social Media based forecasting is that it is a never-ending process – Social Media platforms need to be constantly monitored in order to detect new developments as soon as possible in order to have enough time to take actions.

5.1 Tools for Social Media based forecasting

During this research, diverse tools for Social Media based forecasting were discovered. The big players such as HP, Dell and IBM offer various packages which include Social Media analytics and Statistics Programs comprising Sentiment Analyses, trend spotting, customer analysis and many more functions. For instance, IBM has a wide range of packages on offer which could be of great help during the prediction process, namely *IBM Digital Analytics*, *IBM Social Media Analytics*, *IBM SPSS Data Collection*, *IBM Predictive Customer Intelligence* and many more (IBM, 2015). Additionally, HP offers a *Voice of the Customer* Package which also deals with Social Media Analytics such as text analysis, social media monitoring, fraud and risk mitigation and many more (HP, 2015). Furthermore, several firms offer the process of prediction as a full service. *Trendspottr* for instance, a U.S. start-up company is a platform that predicts forthcoming contents, sentiments, influencers and trends for any subject in any industry using Twitter data and other Social Media channels at real-time. Similarly, *BIG Social Media GmbH*, a German company also offers a number of Social Media software packages including a Social Media Monitoring tool, a Social Media Interaction tool, a Social Media CRM Integration Service, a Social Media Newsroom Editorial Service and a Social Media Tracking & Reporting tool which regularly updates marketers on new developments which affect them on the web. These tools are very valuable for Social Media based forecasting as they reduce workload for marketers and because they comprise everything that is required for successful results.

6. CONCLUSION, LIMITATIONS & FURTHER RESEARCH

In the present paper a clear overview of Social Media based forecasting, its potential of innovative product developments and its accompanying data processing methods as well as limitations and criticisms were outlined. Furthermore a new framework was developed, customized for the prediction of future market trends and customer needs. It can be concluded that predictions can indeed be made with Social Media data, however with caution due to the limitations which were outlined. Various professional tools can be used for support as was outlined in the previous section.

There are several limitations this paper is faced with which have to be considered. The main limitation is the small time frame which this paper had to be completed in. In total approximately ten weeks were available for the entire process i.e. developing an appropriate research problem, collecting, reading and evaluating relevant literature as well as writing out the entire paper. Due to this small time frame it can't be guaranteed that all relevant literature was included in this literature review, perhaps some articles were accidentally missed, also due to restriction issues. Furthermore another restraint which also occurred due to the lack of time is that no empirical research could be conducted which also limits the reliability of this research paper as the new framework could not be tested. Additionally the author could not go into great detail regarding data processing methods due to the restriction of maximum ten pages this paper had to be comprised of. Also, since the topic of Social Media based forecasting especially regarding future trends and customer needs has not reached a stage of maturity, a limited number specific articles could be found. Most articles that were examined during the literature review were quite general or regarding other predictor variables.

Lastly, further research should certainly be conducted regarding the prediction of future market trends and customer needs since this is still quite an untouched topic in the academic field. Empirical research has to be conducted in order to reach a level

of reliability and maturity. Organizations need to be aware of the importance of such forecasting devices which is why this topic needs to be further developed.

7. ACKNOWLEDGMENTS

With this I would like to thank my supervisor Dr. E. Constantinides for his dependable and friendly support throughout this entire Bachelor Thesis research project.

8. REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S., & Liu, B. (2011). Predicting Flu Trends using Twitter data. 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).

Ahn, Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th International Conference on World Wide Web - WWW '07*.

Alexa Ratings. (2015, May 11) Top 500 sites on the web. Retrieved from the Alexa Ratings website: http://www.alexa.com/topsites

Anagnostopoulos, A., Kumar, R., Mahdian, M. (2008). Influence and Correlation in Social Networks. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 7-15.

Asur, S. and Huberman, B. (2010). Predicting the future with Social Media. *IEEE/WIC/ACM, International Conference on Web Intelligence and Intelligent Agent Technology*. 492-499

Bandari, R., Asur, S., Huberman, B.A. (2012). The Pulse of News in Social Media: Forecasting Popularity. *Proceedings on the Sixth International AAAI Conference on Weblogs and Social Media.*

Barbier, G., Liu, H. (2011). Data Mining in Social Media. *Social Network Data Analytics*. Springer, 327-352.

Bendler, J., Brandt, T., Wagner, S., Neumann, D. (2014). Investigating Crime-to-Twitter Relationships in Urban Environments-Facilitating a virtual Neighbourhood Watch. *22nd European Conference on Information Systems.*

Bermingham, A., Smeaton, A.F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP)*, 2–10.

Bifet, A., & Frank, E. (2010). Sentiment Knowledge Discovery in Twitter Streaming Data. *DS'10 Proceedings of the 13th international conference on Discovery Science*, 1-15.

BIG Social Media (June 14, 2015). Products and Services. Retrieved from the BIG Social Media website: http://www.big-social-media.com/software_services/

Bloem, J., van Doorn, M., Duivestein, S., van Manen, T., van Ommeren, E. (2012). Big Social Predicting behaviour with Big Data. SOGETI VINT Research Report 2, retrieved from: http://labs.sogeti.com/download-big-data-reports/

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood Predicts The Stock Market. *Journal of Computational Science 2*, 1-8.

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*

Charalabidis, Y., Loukis, E., Androutsopoulou, A. (2014). Fostering Social Innovation through Multiple Social Media Combinations. *Information Systems Management 31*, 225-239.

Chen, H., Zimbra, D. (2010). AI and Opinion Mining. *Intelligent Systems, IEEE 25*(3), 74-80.

Chung, J., Mustafaraj, E. (2011). Can Collective Sentiment Expressed on Twitter Predict Political Elections? *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 1770-1771.

Constantinides, E., Fountain, S.J. (2008). Web 2.0: Conceptual foundations and marketing issues. *Journal of Direct, Data and Digital Marketing Practice, 9(*3), 231-244.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, 160-168.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, 115-122.

Emerald (2015, May 11). Guide on How to write a Literature Review. Retrieved from: http://www.emeraldgrouppublishing.com/authors/guides/write/literature.htm?part=1

Facebook Newsroom. (2015, May 11) Company Info. Retrieved from the Facebook Newsroom website: http://newsroom.fb.com/company-info/

forecasting. [Def. 1]. (n.d.). In *Business Dictionary*. Retrieved 2015, May 15 from http://www.businessdictionary.com/definition/forecasting.html

Franch, F. (2013). (Wisdom of the Crowds): 2010 UK Election Prediction with Social Media, *Journal of Information Technology & Politics 10* (1), 57-71.

Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management 35*, 137-144.

Gayo-Avello, D. (2011). Don't Turn Social Media Into Another 'Literary Digest' Poll. *Communications of the ACM 54*(10), 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/abs/1204.6441

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, 211-220.

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

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

HP (June 14, 2015). Product brochure: Deliver real-time Voice of Customer Analytics to the contact center. Retrieved from HP website: http://www.hpengage.com/odoc/assets/global/pdf/Products/Promote/Qfiniti/20150603_BR_HP_Explore_web.pdf

IBM (2015). Social Media Analytics – Making Customer Insights Actionable. *Business Analytics*, Retrieved from: http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=YTW03168USEN#

IBM (June 14, 2015) Customer Analytics. Retrieved from the IBM website: http://www-01.ibm.com/software/analytics/rte/an/customer-analytics/products.html

Instagram, Chiara Ferragni (June 14, 2015). Retrieved from Instagram: https://instagram.com/chiaraferragni/

Jahanbakhsh, K., Moon, Y. (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

Kalypso. (2011). Social Media and Product Innovation: Early Adopters Reaping Benefits amidst Challenge and Uncertainty. *Kalypso White Paper*. Retrieved from: http://viewpoints.kalypso.com/uploads/files/Kalypso_Social_Media_and_Product_Innovation_1.pdf

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

Kaplan, A.M., Haenlein, M. (2010). Users of the world, unite! The challenges and oppurtunities of Social Media. *Business Horizons, 53*, 59-68.

Kim, Y., Jeong, S.R. (2015). Opinion Mining Methodology for Social Media Analytics. *KSII Transactions on Internet and Information Systems 9*(1), 391-406.

Lassen, N., Madsen, R., & Vatrapu, R. (2014). Predicting iPhone Sales from iPhone Tweets. *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, 81-90.

Lazer, D., Pentland, A., Adamic, L. et al. (2009). Life in the Network: the coming age of computational social sciences. *Science*, *323*(5915), 721-723.

Liu, Y., Huang, X., An, A., & Yu, X. (2001). ARSA: A Sentiment-Aware Model for Predicting Sales Performance using Blogs.. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 607-614.

Mejova, Y., Srinivasan, P., & Boynton, B. (2013). GOP primary season on Twitter. *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining - WSDM '13*, 517-526.

Metaxas, P.T., Mustafaraj,E. (2012). Social Media and the Elections. *Science and Society 338*, p. 472-473.

Metaxas, P.T., Mustafaraj, E., & Gayo-Avello, D. (2011). How (Not) to Predict Elections. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 165-171.

Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems 13*, 177-195.

Nguyen, L., Wu, P., Chan, W., Peng, W., & Zhang, Y. (2012). Predicting collective sentiment dynamics from time-series social media. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12.* Article no. 6.

Nielsen Norman Group (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 2015, June 2.

O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 1-8.

Oghina, A., Breuss, M., Tsagkias, M., & Rijke, M. (2012). Predicting IMDB Movie Ratings Using Social Media. *ECIR'12 Proceedings of the 34th European conference on Advances in Information Retrieval Lecture Notes in Computer Science*, 503-507.

Plessis, M.D. (2007). The role of knowledge management innovation. *Journal of Knowledge Management 11*(4), 20-29.

Sang, E.T.K., Bos, J. (2012). Predicting the 2011 Dutch Senate Election Results with Twitter. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 53–60.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., Gloor, P. (2013). The Power of Prediction with Social Media. *Internet Research 23*(5), 528-543.

Shadish, W.R., Cook, T.D., Campbell, D.T. (2002). Experimental and Quasi Experimental Designs, Houghton Mifflin Company, New York.

Statista (2015, May 13) U.S. Social Media visit share. Retrieved from Statista website: http://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/

Statistic Brain (2015, May 11) Twitter Statistics. Retrieved from Statistic Brain website: http://www.statisticbrain.com/twitter-statistics/

StatSoft (2015, May 20). What is Data Mining. Retrieved from StatSoft website: http://www.statsoft.com/Textbook/Data-Mining-Techniques#pdm

Thiel, K., Kötter, T., Berthold, M., Silipo, R., Winters, P. (2012). Creating Usable Customer Intelligence from Social Media Data: Network Analytics meets Text Mining. *KNIME*, 1-18. Retrieved from: https://www.knime.org/white-papers

Trendspottr (June 14, 2015) About us. Retrieved from the Trendspottr website: http://www.trendspottr.com/aboutus.php

Truvé, S. (2011). Big Data for the future: Unlocking the Predictive Power of the Web. *Recorded Future, Inc.*

Tuarob, S., Tucker, C.S. (2013). Fad or here to stay: Predicting Product Market Adoption and Longevity Using large scale, Social Media Data. *Proceedings of the ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, 1-13.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity, and other Methodological Pitfalls. *In ICWSM '14: Proceedings of the 8th international AAAI Conference on Weblogs and Social Media*, 505-514.

Twitter. (2015, May 11). About Company. Retrieved from the Twitter website: https://about.twitter.com/company

Wang, X., Gerber, M., & Brown, D. (2012). Automatic Crime Prediction Using Events Extracted from Twitter Posts. *SBP'12 Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 231-238.

Wijnhoven, F., Bloemen, O. (2014). External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? *Decision Support Systems 59*, 262-273.

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

Zeng, D., Chen, H., Lusch, R., Li, SH (2010). Social Media Analytics and Intelligence. *Intelligent Systems, IEEE 25* (6), 13-16.

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*, 55-62.