Social Media as a Source of Predictive Power to Forecast Market Needs

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

h.bockstette@student.utwente.com

The purpose of the following literature review is to verify if social media is a source of predictive power. Various techniques are described along with their different methods to make predictions on future market trends. A broad range of techniques are described and used by the visualized social media analytic process, containing several steps (e.g. capture, understand and present) to predict future market needs derived from Twitter raw-data. The research provides evidence that Twitter is a source to gain information on future customer behavior. Whereas, it concludes, also that techniques and methods needs to be modified to reflect useful information to predict future events. To make efficient use of Twitter raw-data the author concludes with a guideline, namely the "Extended Market Pull Innovation Process", which could be seen as a useful tool for marketing departments and practitioners to establish accurate predictions on market needs, resulting into new innovations.

Supervisors:

Supervisor: Dr. Efthymios Constantinides

Second Supervisor: Dr. M.L. Ehrenhard

Keywords:

Market needs, Innovation, social media analytics, social media and Twitter

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

1. INTRODUCTION

1.1 Social media as a powerful tool to identify market trends, customer behavior and customer needs

The 21st century, predominately provides evidence that social media has evolved into a modern form of traditional gossip. Users post tweets on Twitter to discuss a wide variety of topics including: items people desire to purchase, the quality of a particular service, a product's usability, evaluation of a product, a particular need, and so forth. Significantly a large amount of information is shared between users, which can be obtained by organizations from social media sites including Twitter to support them with a greater amount of useful insights and knowledge on specific customer behavior and certain needs This customer feedback is also known as "consumer insight" (Stone et al., 2004). These information mostly reflect what a consumer is feeling, thinking based on his or her perceptions towards a certain product or brand (Chamlerwatet et al.,2012). This information could give business organizations an opportunity to forecast future customer needs, wants and demands as well as market requirements, in short also know as market needs. By exploiting these data companies may have an advantage to enhance their innovative capabilities. The current thesis provides recommendations for organizations to exploit these information to fulfill market needs, which ultimately leads to better innovations, higher profits and efficient operations. Recent academia provides evidence that the identification, collection and analysis of social media data is essential for a manager's ability to make adequate decisions within an organization (Harald, et al. 2013). Similarly, Wolfers and Zitzewitz (2004) have shown that social media is indeed an effective medium to make predictions. Over many years business organizations have been challenged when they are faced with unexpected market needs, due to rapidly changing trends. Existing opportunities and the potential could not be fully explored and therefore R&D efforts didn't succeed to produce innovative products/services to boost profits. However, it is sometimes the case that innovations "come out of the blue" or "fall from the sky", but in most cases observing and analyzing the market increases the success rate of product/service innovations. Therefore, a prediction of market needs and trends would give companies a competitive advantage and enable them to produce innovations that costumers value and willing to pay for. A complaint on Twitter about particular problems or lack of functionality can be seen as a feature request, which indicates a specific market need and may lead to innovation. This paper will support not only marketing departments but also practitioners in identifying useful data from social media to forecast or predict future events. The paper serve as a guideline to facilitate managers in their decision making process by explaining how data from social media could be transformed into successful innovative applications.

1.2 Key terms

The key terms are defined to give the reader a general overview, namely: Market needs, Innovation, social media analytics, social media and Twitter

1.2.1 Market needs

According to the online dictionary "Businessdictionary.com" market needs could be defined as " *A driver of human action which marketers try to identify, emphasize, and satisfy, and around which promotional efforts are organized*". A more general definition of a market need is " *a motivating force that compels action for its satisfaction*". So a market need in the context of the current thesis is something a big group of people (market) wants to have or needs to have (need). Hence, the notion Market needs in this thesis comprises the focus on consumer needs, wants and demands as well as market requirements

1.2.2 Innovation

An innovation (radical or incremental) occurs if an entity or a person makes a significant contribution to a product/service that is already known and is one of the primary ways to differentiate your product/service from the competition. So it is not just about designing a new product, service or process but also improving existing ones (e.g. Improve efficiency or cut down waste). So if strong price competition exists, there is a need for innovations to gain a competitive advantage. According to Schumpeter, (1934) innovation can be regarded as an invention, which is commercialized and brought to the market by entrepreneurs. However, let us start from the beginning. Everything begins with an idea, which is neither an invention nor an innovation. However, ideas could lead to inventions when it is converted into tangible new artifact (Trott, 2005). This transformation is called the invention process. The innovation process illustrated in figure 1, consist of the invention process and subsequent operations build upon the invention for making commercial success (Trott, 2005). Schoen et al. (2005) see this process as a complex, non-linear, iterative process, which includes elements of randomness.

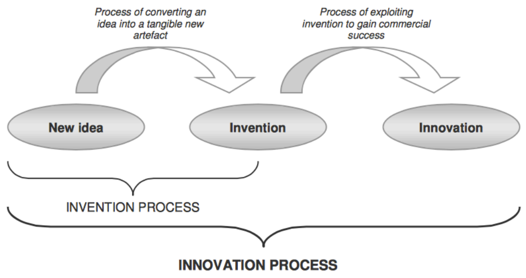

Process of converting an idea into a tangible new artefact
Process of exploiting invention to gain commercial success
New idea
Invention
Innovation
INVENTION PROCESS
INNOVATION PROCESS

Figure 1: Innovation Process (Trott, 2005)

Schumpeter (1934) traditionally classified innovations into two different types: product and process innovations. Trott (2005) adds up five new classifications namely: organizational, management, production, commercial and service innovations next to the conventional: Product and process innovations. Tidd et al. (1998) divide innovation into the degree of novelty. He speaks about incremental and radical innovations. An incremental innovation is to improve existing products/services, leading to small improvements On the contrary a radical innovation is a fundamental change in products, services or processes. However, academia provides evidence that a radical innovation is rare in is occurrence (O'connor et al. 2002). Rothwell (1992) has developed five different generations of the innovation process to produce innovations. They reach from simple linear to complex interactive models. However, the focus of this Thesis lies on the Market Pull model (see Figure 2). The "Market pull" strategy is a simple linear sequential process with a focus on marketing. This model views the market as a source of new ideas for the R&D activities. In that model consumer needs, wants and demands as well as market requirements, summarized as market needs, are seen as the key driver for innovation. It includes and integrates user conceptions in the innovational process that could be gathered by using social media content.

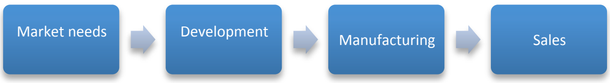


Figure 2: Second Generation of the innovation process: Need Pull

1.2.3 Social media analytics:

Social media analytics deals with the measurement, analysis and interpretation of the interactions and associations between people's topics and ideas. According to Zeng et al., (2010) it is an anxious of creating and evaluating tools and frameworks in order to collect, monitor, analyze, summarize, and visualize social media data with specific requirements from a target application. It gives researchers the ability to uncover for example customer sentiments, which can provide further information and insights concerning market needs.

1.2.4 Social media

Kietzman et al. (2011) defines social media as "*mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, cooperate, discuss and modify user-generated content*" (p. 241). According to Garner (2013), it is expected, that worldwide corporate organizations and institutions will make use of social media as a primary source to communicate with their clients and stakeholders: both internal and external. Different types of social media are presented by Kaplan and Haenlein (2009; 2010; 2011) namely: micro blogs such as Twitter, content communities like Youtube or Instagram, social networks like Facebook and collective projects like Wikipedia. The importance of social media platforms increased in the last decade since they offer incremental changes in the way organizations and individuals communicate with each other. (Kietzman et al., 2011). The importance of social media and especially media marketing strategy for companies, regardless of the size and industry, convey since Hahn, Rohm and Critenden (2011) revealed, that social media has transformed the internet " *from a platform for information to a platform for influence*" (p. 272). Users spend over 20% of their time online on social media platforms (Fan & Gordon, 2014). Social media is important, because it offers benefits to businesses and entrepreneurs leading to minimal costs. Businesses have to ensure, that their social media platforms are interactive and competitive, when benchmarked with other industrial players. When conducting marketing to increase growth and brand awareness for different organizations and companies, it is important to use social media as a measure to forecast future events (Agnihotri, Kothandaraman, Kashyap, Rajiv & Ramendra. 2012). When it comes to predictions, Twitter needs to present more accurate and detailed information, since academia provides evidence that information from Twitter is used already as the basis for many predictions made by other literature e.g. disease outbreak or influenza (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Culotta, 2010; Ritterman, Osborne,& Klein, 2009; Signorini, Segre, & Polgreen, 2011), election outcomes (Lui et al,. 2011; Franch, 2013; Sang &Bos, 2012; Mataxas, Mustafaraj, & Gayo-Avello,2011: Tumasjan, Sprenger, Sandner, & Welpe, 2010) and also sales and Revenues (Lassen et al., 2014, Asur & Huberman, 2010). Facebook can be neglected even though it is one of the biggest social media networks around. According to Couper (2013), messages, comments and posts is not publicity available and hence not easy to collect. Therefore research will focus solely on the presented social media platform Twitter:

1.2.5 Twitter

Twitter, as a sub form of social media, is a form of a microblog. The goal of Twitter is to spread information around the world (Cha, Benevenuto, Haddadi & Gummadi, 2012). To use Twitter, an account is necessary, either personal or anonymous, using a personal fake information's. Like already stated, the core intention of Twitter is to spread information, either in form of real information, rumors or own opinions. The informational content can be seen by users' followers or even by all Twitter account members, who are involved in the certain topic discussion. To determine which topic the author wishes to address in his post, the author needs to put a hash tag (#) symbol in front of his respective keywords and the post will be part of the addressed issue, which is discussed by other users in the community using the same hash tag (Romero, Meeder, & Kleinberg, 2011). Besides private participants, everybody can "follow" sites of well-known institutions. The author used Twitter as an accurate source of Data for predicting market needs since it exhibits according to Culotta (2010) some important attributes.

a. "The high message posting frequency enables up-to-the-minute analysis (...)"

b. "As opposed to search engine query logs, Twitter messages are now longer, more descriptive, and (in many cases) publically available."

c. "Twitter profiles often contain semi-structured meta-data (city, state, gender, age), enabling a detailed demographic analysis."

d. "Despite the fact that Twitter appears to target a young demographic, it in fact has quite a diverse set of users. The majority of Twitter's nearly 10 million unique visitors in February 209 were 35 years or older, and a nearly equal percentage of users are between age 55 and 63 as are between 18 and 24."

The purpose of this research is to review present findings on social media analytics to gain insights on techniques and tools to gather and identify market needs in the first place and innovations by extension.

Hence, the research problem is stated as follow: "To what extend can companies use Twitter to predict market needs?"

In order to propose an accurate conclusion to the research problem, the following research tries to answer sub-research questions, which are namely: (1) What is the social media analytics process (2) What are the main social media analytic techniques and their methods used in the analytic process?

2. METHODOLOGY

The information provided within the critical literature review are rested on former academic studies, which are retrieved from Web of Science, Google Scholar and by searching through several other journals (MISQuarterly, Information System Research and the Journal of MIS). The thesis, especially focus on the most recent literature, starting from year 2008. The primary keywords to search for potential literature were " social media forecasting", "Forecasting with social media" and "Predictive power of social media". After selecting and analyzing some articles, keywords like "Twitter Predictions" and "Twitter forecasting" have been used more frequently because these social media platform were often used to make predictions on specific topics. More contributing articles have been selected by checking the reference lists of former selected articles, which resulted in a sum of 70 articles. Acceptable papers regarding innovations have been found by searching for key terms including "innovation", "innovation process" and "Steps of innovations". By exploring the reference list, more articles have been found. For reviewing the different methods used in each technique, the author searched for terms such as "Sentiment analysis methods", "Visual analysis methods". Further, searching for keywords within different articles derives the literature. The research questions, mentioned in section 1, can be seen as a guideline for the critical review in chapter 3, which will provide the reader with a review of the current findings on social media analytics. Finally, this thesis ends with a conclusion in Section 4, by describing the author findings concerning the research problem and a visualized extended model of the market pull innovation process.

3. CRITICAL LITERATURE REVIEW

The development of knowledge on future market needs is essential for companies to capture opportunities for breakthrough innovations. Especially the process of social media analytic with aligned techniques and methods support companies in acquiring information related to market needs. Therefore, the following chapter will introduce the analytic process and used techniques.

3.1 The Process of Social Media Analytics

According to Fan and Gordon (2014) the social media analytics process consists of three stages: (1) Capture, (2) Understanding and (3) Present (see Figure 4). The Capture stage tries to explore relevant social media data by observing, monitoring and "listening" to social media platforms. Major tasks include recording data and extracting useful information by using APIs (Application Performance Interfaces) or through crawling applications. Especially, API's supportive for and accommodative to personal needs can be found online[1]. The company itself could gather crawling data from third-party vendors. However, not all captured data can be used within the process of social media analytics. In order to establish a adequate dataset for upcoming stages, several preprocessing steps need to be performed, such as data modeling, and record linking, stemming, part-of-speech tagging, feature extraction, and other syntactic and semantic operations that support the process of analytics (Fan & Gordon, 2014). The second step is Understanding, which considers the selection of relevant data, while removing noisy low-quality data. This is majorly done by using different advanced data analytic methods. Hence, Understanding is about gaining insights by exploring data through statistical methods, which are especially data mining, natural language processing, machine translation, and network analysis, meaning and generating useful metrics for decision making (Fan, 2006). Important to note in this context is that the process of Understanding is seen as a crucial stage in performing social media analytics, since the results will significantly affect the information and metrics used in the final Present stage. This stage reveals the findings and outcomes from the two former steps. It is similarly a summary of key-findings presented in an easy-to-understand format, basically with the support of visualization techniques and data-statistical analysts serving as decision makers, in creating important information.

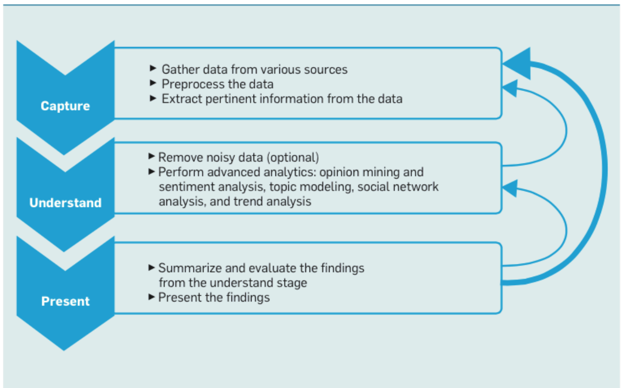

Capture
▸ Gather data from various sources
▸ Preprocess the data
▸ Extract pertinent information from the data
Understand
▸ Remove noisy data (optional)
▸ Perform advanced analytics: opinion mining and sentiment analysis, topic modeling, social network analysis, and trend analysis
Present
▸ Summarize and evaluate the findings from the understand stage
▸ Present the findings

Figure 3: Based on Gordon & Fan (2014) Social Media Analytics Process - The CUP Framework

3.2 Social media analytic techniques

Social media analytics covers various modeling and analytical techniques derived from different fields. The following part will highlight specific instrumental techniques mostly used in social media analytics. Gordan and Fan (2014) build the foundation with the CUP-Framework, these stages are predominately filled with different techniques. For the Understanding stage, the techniques are - topic modeling and social network analysis, however, they have primarily been used as an application. Other techniques used in the Understanding stage include sentiment analysis and trend analysis. Whereby, visual analytics spans the second and third phase of the social media analytic process. For a better overview see figure 5.

[1] http://help.sentiment140.com/other-resources

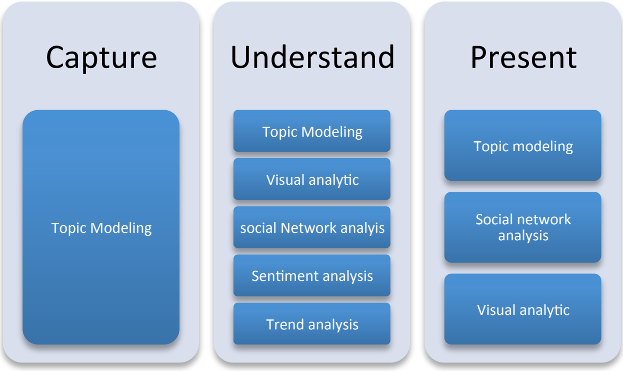
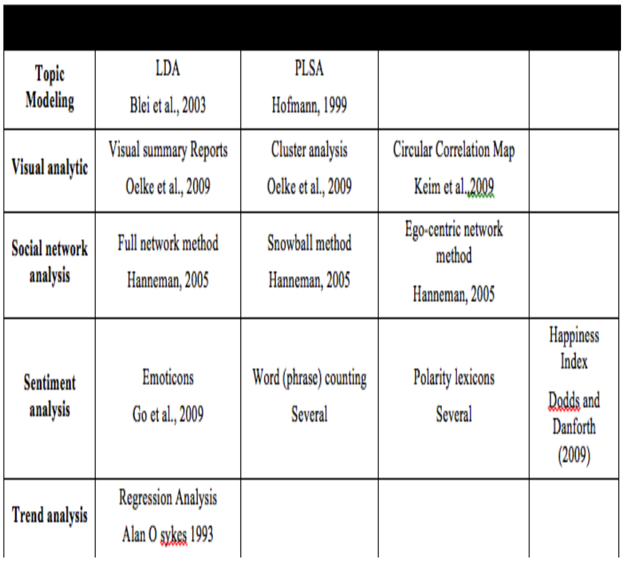


Figure 4: Techniques used in the various steps of the social media analytic process

In the following sub-sections the different techniques and incidental methods will be presented (Table 1):



Topic Modeling	LDA Blei et al., 2003	PLSA Hofmann, 1999		
Visual analytic	Visual summary Reports Oelke et al., 2009	Cluster analysis Oelke et al., 2009	Circular Correlation Map Keim et al.,2009	
Social network analysis	Full network method Hanneman, 2005	Snowball method Hanneman, 2005	Ego-centric network method Hanneman, 2005	
Sentiment analysis	Emoticons Go et al., 2009	Word (phrase) counting Several	Polarity lexicons Several	Happiness Index Dodds and Danforth (2009)
Trend analysis	Regression Analysis Alan O sykes 1993			

Table 1: Categorization of different techniques based on specific methods

3.2.1 Topic Modeling

Topic modeling is defined as the detailed gathering for dominant topics or themes by browsing through captured text. As already mentioned in the previous section, this technique is primarily used in the Understanding stage, but can also support capture and present stage. According to Blei (2012), topic models are " (…) *algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes*". (…) can be applied to "*massive collections of documents*" and can be "*adapted to many kinds of data*" (Blei, 2012) such as Twitter data. Topic modeling is " (…) *a way of identifying patterns in a corpus*" (Brett, 2012). Similarly, Posner (2012) argues that topic modeling is " *a method for finding and tracking clusters of words (called "topics" in shorthand) in large bodies of texts*". From a general perspective, topic modeling is like going through a text with different colored highlighters by assigning each keyword a different color. All keywords marked with the same color belong to the same group, and each color represents a different topic. There a various advanced statistics and machine-learning techniques for performing topic modeling. For instance, Hofmann (1992) illustrates models that identify "latent" topics through "*the co-occurrence frequencies of word within a single communication*" and Yin et al. (2012) has observed a relation between topics and communities of users. An unsupervised machine learning approach, which is designed to identify latent topic information in large texts is **LDA (Latend Dirichlet Allocation).** It makes use of the so-called "bag of words" approach and treats every document as a vector of word counts (Hong & Davison, 2010). This method does not provide explicit information about authors' interests but delivers additional information on content of documents. For an exact work routine Hong and Davison (2010) state that: "*Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words*". LDA generate a document within a three-step process: "First, for each document, a distribution over topics is sampled from a Dirichlet distribution. Second, for each word in the document, a single topic is chosen according to this distribution. Finally, each word is sampled from a multinomial distribution over words specific to the sample topic" (Rosen-Zvi et al., 2004). This process is based on the Bayesian model. For a detailed explanation see Blei et al., 2003. Another method is **PLSA** (Probabilistic Latent Semantic Analysis**)** that is according to Hofman (1999) a "*novel approach to automated document indexing that is based on a statistical latent class model for factor analysis of count data*". This statistical technique is a method to analyze two-mode and co-occurrence data by using multinomial distribution. For a complete overview see Hofman (1992). However, explaining these techniques in detail would expand this thesis therefore the author point to Blei (2012), Brett (2012), Hofmann (1992) and Yin et al. (2012) for a detailed overview on topic modeling.

3.2.2 Visual analytics

Visual analytics supports the "Present" step of the social media analytics process. According to Thomas & Cook (2006) visual analytics is "*the science of analytical reasoning facilitated by interactive visual interfaces*". Basically developed to fulfill the needs of the U.S. defense force. Visualization can be used as an enhancer in different situations such as synthesis, exploration, discovery and confirmation to get insight from raw data that is mostly voluminous and derived from various sources. Additionally, Visual analytics enables data collection leading to data-supported decision-making procedures. Therefore, many different statistical models are build on the foundation of visual analytics, whereby the human ability to recognize patterns and draw conclusions is seen as a key factor within the process as a whole. It is important to connect human strength and machines to make accurate decisions, which are well explained and justified. A more user perceptual view, is to see visual analytics as "*a collection of techniques that use graphical interfaces to present summarized, heterogeneous information that helps users visually inspect and understand the results of underlying computational process*" (Fan & Gordon, 2014). As stated earlier, visual analytic is the process of data visualization to present an easy access. Oelke et al., 2009 describes three different methods (1) Visual summary reports (2) Cluster Analysis and (3) Circular correlation map. They certainly use these methods in a different context than Twitter. However, it should also work properly with Twitter data as well. **Visual summary reports** method may be considered to provide a quick overview on customer feedback data sets, eliminating extensive reading. An automatic algorithm marks each attribute either blue if it's a more positive feedback and red, if the feedback is more negative. Size of the inner rectangle (see figure 6) reflects the "*percentage of reviews that commented on the attribute signaling the importance that the analyst should give to this attribute in his or her evaluation*" (Oelke et al., 2009).

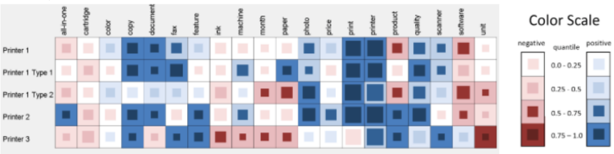


Figure 5: Performance of Printers (Oelke et al., 2009).

A **cluster analysis** tends to play a key role for companies to access and analyze different customer groups based on similar opinions. (Oelke et al., 2009). Keim et al., (2006) introduced the **Circular Correlation map** to offer a detailed view on specific data. It supports managers and practitioners in detecting correlations between the different variables of the whole data set (Oelke et al., 2009). Cluster analysis has the advantage of detecting trends, correlations or patterns from data (Ananiadou, 2008). See Keim et al., (2006) and Oelke et al., (2009) for a more detailed description and details on the work routine.

3.2.3 Social Network analysis

Social network analysis could be defined as major technique to identify key influencers in viral market campaigns on social media, which is based on a social network graph, categorized by nodes (users) and associated relationships (edges) and tries to understand the underlying structure, connections, and theoretical properties. Hence, Social network analysis focuses on the relations among actors. This tool is especially useful in detecting sub-communities within a large online community to establish more precise product tailoring and marketing materials. Bonchi et al., (2012) underlines that social network analysis could enhance predictive modeling. However, methods that have predated the advent of social media and focus more on analyzing static mathematical graphs, needs to be adjusted to continually changing network structures by developing the required new technical approaches. Scott (2012) reinforces this statement in his book: "Social network analysis" and more social media related insights are given by Hanneman (2005). The key terms within the context of Social network analysis could be defined as follow: Nodes: actors such as Twitter user; Ties: relations between actors; Ego: Main subject of study. The first social media analytic method is the **Full network method**. This method needs information about each actor's ties with all other involved (Hanneman, 2005). It counts all ties in a population of actors. Examples are given by Hanneman, (2005) it could be collected "(…) *data on shipments of copper between all pairs of nation-states in the world system from IMF records; we could examine the boards of directors of all public corporations for overlapping directors; we could count the number of vehicles moving between all pairs of cities; we could look at the flows of e-mail between all pairs of employees in a company; we could ask each child in a play group to identify their friends. With this method a full picture of the whole relation can be identified*". These methods are very powerful to give a detailed description and analyzes on social structures. Unfortunately this method takes much time and social media has made this process less complex and to a favorable price. **The snowball method** starts with a specific person or group of interest. As a starting point participants are asked to name their ties, whereby all actors who were named but not were part of the original list are tracked down and asked for some of their ties. This process goes on until no new actor will be identified, or the process will be stopped due to time or costs. (Hanneman, 2005). This method is particularly helpful for identifying a "special" population such as business contact networks, deviant sub-cultures or football club fans. Noteworthy in this context is to mention that the start should be based on the best initial node. Limitations of the method are twofold: the isolation of actors who are not connected e.g. because they do not have Twitter or Facebook and second there is no way to explore all the connected individuals. In many situations, it is not necessary to track down the full networks beginning with a focal node as described in the snowball method (Hanneman, 2005). Therefore, an alternative approach is considered namely, the **egocentric network method,** which has been developed to start with a selection of focal nodes (egos) and the identifications of connections of the belonging nodes (Hanneman, 2005). Afterwards, nodes that are defined in the first stage will be checked if they are connected to one another. Asking each node can do this. This method is suitable for detecting a "(…) *form of relational data from very large populations*, (...)" (Hannemann, 2005) and can give the researcher a reliable picture of existing networks. Exemplary, Hanneman (2005) states " (…) *we might take a simple random sample of male college students and ask them to report who are their close friends, and which of these friends know one another*". This approach can give researcher also additional information about the network as a whole. However, it is not that complete like the snowball or census approach.

3.2.4 Sentiment Analysis

Sentiment analysis is defined as the core procedure behind any social media trend-analysis applications and monitor systems. Analytic text methods are used such as natural language processing or computational linguistic to extract information on social media user sentiments and opinions about specific topics. The extracted information is about people, products or services, which is always subjective, but already used successfully to make various predictions such as stock market movements (Zhang et al., 2010). There is a broad outline of two different types of sentiment methods, based on: machine learning and lexical. Both methods often try to identify a specific polarity, which is a positive or negative attitude towards a particular topic. While machine learning methods often rely on "*supervised classification approaches,* [whereas] *sentiment detection is framed as a binary* [process] *(i.e. positive or negative)"* (Goncalves & Araujo, 2014). Lexical-based methods tend to produce a word list in order to assign every word to a specific sentiment. Both have beneficiaries and disadvantage in their daily use. But the ability to adapt and create sophisticated models for concrete situations is an advantage for the usage of machine-learning-based methods (Goncalves & Araujo, 2014). On the contrary the benefit of the lexical-based methods, which varies according to the context they have been created for, is that they could be adjusted or modified to personal needs. Academia provides evidence that machine-learning methods are more convenient than lexical-based methods (Tausczik & Pennebaker, 2010). Methods of sentiment analysis reach from relatively simple methods to more complicated and sophisticated methods. The easiest method to detect sentiments is by exploring the different types of emoticons in tweets. **Emoticons** primarily reflects positive or negative feelings. Goncalves and Arajuo (2014) developed a table, visualizing various types of Emoticons representing particular feelings a consumer desires to express. Next to this, the **happiness index** developed by Dodds and Danforth (2009) is an additional, relatively easy method to detect polarity. The happiness index makes use of the Affective Norms for English Words (ANEW)

(Bradley & Lang, 1999). Additionally, the index scores and ranks a text between 1-9, indicating happiness or sadness of the text. Within a text the frequency of each wording is calculated to produce an ANEW, which is finally the input for the weighted average of the valence on ANEW study words. Goncalves and Arauo (2014) considered that a text has a negative perception with a score from 1-5 and a positive perception with a score from 5-9. Further methods are identified such as the **word (phrase) counting**, which assumes, for example that the frequency of in product wordings within a text has an influence on customer perception. Another method is to list positive and negative terms in combination with mentioning a product by name that can be extracted from social media content. Positive terms could be stated as follow: "brilliant", "phenomenal", excellent", whereby negative terms could be: "suck", "terrible" or "awful". (Pang & Lee, 2008). This method is known as **polarity lexicons**. Similarly, academia speaks about semantic methods that may compute the lexical "distance" between products name and each of two opposing terms like "good" and "bad" (Turney, 2002). Researchers have developed several tools to make the process for word (phrase) counting and polarity lexicons faster and easier (e.g. LIWC, SentiStrength, SenticNet, and Pansas-T). All these tools have been designed for Twitter applications. LIWC[2] (Linguistic inquiry and Word count) is a text analysis tool to assign emotional words in a given text to their classified categories such as positive or negative. The classification is based on a dictionary that assigns words to a specific sentiment. Additionally, this method can provide other sets of sentiment categories by a given word. For example the word "agree" can be matched to the word categories: assent, affective, positive emotion, positive feeling and cognitive process (Goncalves & Arajuo, 2014). LIWC is a software that can be bought. The software offers users the opportunity to explore customized dictionaries next to the standard ones. SentiStrength software is a dictionary-based attempt, which is developed to extract simultaneously positive and negative sentiments from a given text. This requires that SentiStrenght is adjusted to the writing style and grammatical annotations of social media texts (Thewall et al., 2010). The input is thereby a short electronic text that leads to an evaluated output, displaying expressed sentiments with both: strength of positive and negative feelings. The next method is the SentiWordNet[3] (Esuli and Sebastiani, 2006) tool, based on an English lexical dictionary called wordNet (Miller, 1995) and mainly used in opinion mining. WordNet builds synonyms called "synsets" by grouping adjectives, nouns, verbs and other grammatical classes (Goncalvea & Araujo, 2014) to indicate the sentiment of the text by scoring these synsets. Synsets can be positive, negative or neutral, and the score values range from 0 to 1 and will be aggregate up to 1. For example, a given synsets of a text will be s=(bad, wicked, terrible) the given score for the synsets is positive=0, negative= 0.850 and neutral=0.150 (Goncalves & Arajo, 2014). A semi-supervised machine learning method will evaluate scores. A further software/tool for sentiment analysis and opinion mining is SenticNet[4], which originally was considered as a tool to measure the polarity in opinions of patients in England (Cambria et al., 2010). It explores not only semantic web techniques but also artificial intelligence. The purpose of SenticNet is to evaluate if a common sense concept from natural language text (NLP) is more positive tempered or negative on a semantic level. Goncalves and Araujo give an example of the work routine of SenticNet, "The method uses Natural Language Processing (NLP) techniques to create a polarity for nearly 14,000 concepts. For instance, to interpret a message "Boring, it is Monday morning", SenticNet first tries to identify concepts, which are "boring" and "Monday morning" in this case. Then it gives polarity score to each concept, in this case, -0.383 for "boring", and +0.228 for "Monday morning". The resulting sentiment score of SenticNet for this example is -0.077, which is the average of these values. "PANAS-t (Goncalves et al., 2013) method's purpose is to track any increase or decrease in sentiments over time. It is based on the Positive Affect Negative Affect scale (PANSAS) by Wattson and Clark (1985). Pansas-t consists of word associated with eleven moods: joviality, assurance, serenity, surprise, fear, sadness, guilt, hostility, shyness, fatigue, and attentiveness. To connect the text to a specific sentiment, PANSAS-t makes use of the normative values of each sentiment base on the complete data. Afterward, it computes the P(s) score, range from -1 to 1 for each sentiment within a specific time period to indicate the change. Goncalves and Arauo, 2014 are giving a good example: "a given set of tweets contain P ("surprise") as 0.250, then sentiments related to "surprise" increased by 25% compared to a typical day. Similarly, P (s) = −0.015 means that the sentiments decreased by 1.5% compared to a typical day."

3.2.5 Trend Analysis

Trend analysis is predominantly based on historical data, which is collected over time. It is based on statistical methods like regression analysis (Anderson, 1971). Trend analysis is frequently used to forecast growth of customer, sales numbers, consumer sentiments, the effectiveness of marketing campaigns and stock market movement. If a relationship between two variables, the dependent and independent one, needs to investigate, **regression analysis** is a useful method. With it a causal effect of one variable upon another can be identified (Sykes, 1993) (e.g. the impact of a price increases upon demand). Hence, regression analysis can be an excellent method to make predictions and future forecasts. Former literature already made use of regression analysis to predict the future using Twitter data (e.g. Asur & Huberman, 2010) made use of regression analysis to predict box-office revenue for movies and Lassen et al. (2014) have used regression analysis for predicting iPhone sales. A regression analysis is capable to model a relationship between several independent variables and/or a curvilinear relationship. However, the forecasting ability heavily depends on the accuracy of the estimates for the independent variable and a consistent relationship between variables is assumed, which is not always the case. Of course, there are other statistical methods to perform a trend analysis (e.g. ANOVA method, neural network analysis, vector machines). But, Regression analysis is the most common method used in Trend analysis

4. CONCLUSION

The purpose of this research was to find an answer to the following research question: "To what extent can companies create innovations using Twitter". At first it was necessary to define several key terms such as market needs, Innovation, social media analysis, social media, analytic social media and Twitter to give the reader an overview of important terms and concepts. To provide a valid answer, the author has developed a set of research questions that are ordered in a consecutive way.

[2] http://liwc.net/.
[3] http://sentiwordnet.isti.cnr.it/.
[4] http://Sentic.net/.

Reinforcing the innovation process model (see Figure 2) the first step is to find specific information's on market needs based on customer insights, which is necessary to create innovative applications. Of course, there are other strategies for creating innovations such as the Market Push strategy. However, to cope with the challenge of producing adequate information, data needs to be analyzed and evaluated. Analyzing is a broad notation that includes a set of different steps. Therefore, the author reviewed the social media analytic process to give establish a concrete overview on how to analyze social media data, which consequently results in new ideas for innovative applications to fulfill future market needs. As illustrated in this thesis, Twitter is a good source to gather information on users' insights. Of course natural obstacles in using Twitter exist, because situations may exist where satisfied users do not tweet but unsatisfied ones loudly express their opinions. Twitter raw data is useless and needs to be transformed into useful information by passing the specific information through the social media analytic process. The social media analytic process consists of three interdependent steps. In step 1 researcher "captures" the data, in step 2 researchers tries to "understand" the data and in step 3 researchers "present" it. Within the social media analytic process, different techniques for each step are in visualized (see figure 4), and each technique consists of different methods (see table 1). To get a broad overview of techniques and its methods used in the social media analytic process, the author stated the second research question "What are the main social media analytic techniques and their methods used in the analysis process?" Most cited techniques, and associated methods have been presented. In the author's opinion the understanding stage is the most important step to examine market needs that are foremost a qualitative aspect. As defined in the introduction, market need is "*something a big group of people (market) wants to have or needs to have (need)"*. Therefore we need to identify customer insights to perform accurate market need predictions. Rungkasiri & Haruechaiyasak (2012) find out that "*sentiment analysis on micro-blogs is a useful tool for the consumer research*" because "*a wide range of human moods can be captured through sentiment analysis*" (Goncalves & Arauo, 2014) Therefore to get customer insights, the sentiment technique would be appropriate. For example, if an electric-car producer notices a large number of defaults on the durability of the battery, the usage is connected with a negative sentiment, they can predict a demand for a more durable battery and start the innovation process to create better batteries for their cars.

Reinforcing the research problem "To what extend can companies use Twitter to predict market needs? " it can be said that no innovation will fall out of the blue analyzing Twitter content, however, Twitter can serve as a pool for information on market needs by going through the social media analytic process. With that knowledge of future market needs companies can derive new ideas for the process of innovation. Needless to say is that the presented techniques and methods need to be adjusted to the conditions and the prevailing data structure of Twitter. In Figure 6 the author presents an extended model of the market pull innovation process in which it comes clear how Twitter data leads to innovations. Therefore, companies using a market pull strategy to create innovations can use Twitter content to predict market needs, which in turn can be used to create innovation. Also Twitter makes the innovation process cheaper, easy and lead to innovations that are needed or demanded by the people.

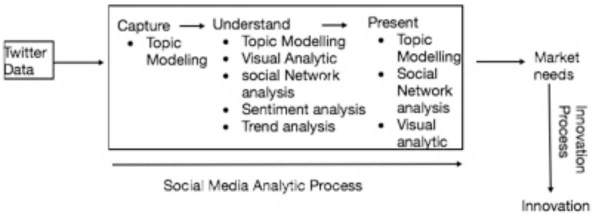


Figure 6: The Extended Market Pull Innovation Process

4. FURTHER RESEARCH

This research has especially focused on the innovation market pull approach. Future researcher could try to explore the relation between innovation (e.g. technology) push approach and social media data (Twitter) and how this relation may affect the possibility or opportunity to predict future customer behavior. Empirical research could be considered to test the degree of innovative capability of companies using especially Twitter as a source for R&D activities. These results could be gathered and measured to verify if Twitter has a positive effect on innovative activities. The Extended Market Pull Innovation Process builds a foundation to include also additional social media data, which should be taken into consideration for further research.

5. ACKNOWLEDGEMENT

I would like to express my very great appreciation to dr. Efthymios Constantinides for his effort during my Bachelor thesis by giving outstanding feedback and advice. I am also appreciative for his assistance and support in focusing on milestones and deadlines. .

6. REFERENECES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011, April). Predicting flu trends using Twitter data. In Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on (pp. 702-707). IEEE.

Agnihotri, R., Kothandaraman, P., Kashyap, R., & Singh, R. (2012). Bringing “social” into sales: The impact of salespeople’s social media use on service behaviors and value creation. Journal of Personal Selling & Sales Management,32(3), 333-348.

Anderson, T.W. *The Statistical Analysis of Time Series*. John Wiley & Sons, Inc., New York, 1971.

Ananiadou, S. (2008). National centre for text mining: Introduction to tools for researchers. Retrieved from http://www.jisc.ac.uk/publications/publications/ bpnationalcentrefortextminingv1.aspx Accessed 08.02.2009.

Asuncion, H. U., Asuncion, A. U., & Taylor, R. N. (2010, May). Software traceability with topic modeling. In Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1 (pp. 95-104). ACM.

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.

Bermingham. A and Smeaton A.F. Classifying Sentiment in Microblogs: Is Brevity an Advantage? In ACM International Conference on Information and Knowledge Management (CIKM), pages 1833–1836, 2010.

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Blei, D.M. Probabilistic topic models. *Commun. ACM 55*, 4 (Apr. 2012), 77–84.

Bradley.M.M and Lang.P.J. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999

Brett,M. (2012). Topic Modeling: A basic Introduction. Journal of Digital Humanities

Bonchi, F., Castillo, C., Gionis, A., and Jaimes, A. Social network analysis and mining for business applications. *ACM Transactions on Intelligent Systems and Technology 2*, 3 (Apr. 2011), 22.

Cambria, E., Hussain, A., Havasi, C., Eckl, C., & Munro, J. (2010). Towards crowd validation of the UK national health service. WebSci10, 1-5.

Cha, M., Benevenuto, F., Haddadi, H., & Gummadi, K. (2012). The world of connections and information flow in Twitter. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42(4), 991-998.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one, 5(11), e14118.

Choi, H., & Varian, H. (2012). Predicting the present with google trends.Economic Record, 88(s1), 2-9.

Couper, M. (2013, January). Is the sky falling? New technology, changing media, and the future of surveys. In Survey Research Methods (Vol. 7, No. 3, pp. 145-156).

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. arXiv preprint arXiv:1007.4748.

Dodds,P.S and Danforth, C.M. Measuring the happiness of large-scale written expression: songs, blogs, and presidents. Journal of Happiness Studies, 11(4):441–456, 2009.

Esuli and Sebastiani. Sentwordnet: A publicly available lexical resource for opinion mining. In International Conference on Language Resources and Evaluation (LREC), pages 417–422, 2006.

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. Communications of the ACM, 57(6), 74-81.

Fan, W., Wallace, L., Rich, S., and Zhang, Z. Tapping the power of text mining. *Commun. ACM 49*, 9 (Sept. 2006), 77–82.

Franch, F. (2013). (Wisdom of the crowds) 2: 2010 UK election prediction with social media. Journal of Information Technology & Politics, 10(1), 57-71.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

Grossberg, S. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks 1*, 1 (1988), 17–61.

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods.

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

Hofmann, T. Probabilistic latent semantic indexing.In *Proceedings of the ACM SIGIR International Conference on*

Research and Development in Information Retrieval (Berkeley, CA). ACM Press, New York, 1999, 50–57.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in Twitter. In Proceedings of the First Workshop on social media Analytics (pp. 80-88). ACM.

Inouye, D., Ravikumar, P., & Dhillon, I. (2014). Admixture of Poisson MRFs: A topic model with word dependencies. In Proceedings of The 31st International Conference on Machine Learning (pp. 683-691).

Johnson, D. (2001). What is innovation and entrepreneurship? Lessons for larger organisations. Industrial and Commercial Training, 33(4), 135-140.

Kaplan, A. M., & Haenlein, M. (2011). Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. Business Horizons, 54(3), 253-263.

Keim, D. A., Hao, M. C. Dayal, U., Janetzko, H., and Bak, P., Generalized Scatter Plots. Information Visualization Journal (IVS09)

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). social media? Get serious! Understanding the functional building blocks of social media. Business horizons, 54(3), 241-251.

Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector-tracking epidemics on Twitter. In Machine Learning and Knowledge Discovery in Databases (pp. 599-602). Springer Berlin Heidelberg.

Livne, M. Simmons, E. Adar, and L. Adamic, "The party is over here: Structure and content in the 2010 election," in Proc. of 5th ICWSM, 2011. [Online]. Available: http://bit.ly/q9lSug

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the US elections through search volume activity. Paper presented at the Proceedings of the IADIS International Conference on e-Society.

G. A. Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series.ICWSM, 11, 122-129.

Oelke, D., Hao, M., Rohrdantz, C., Keim, D. A., Dayal, U., Haug, L., & Janetzko, H. (2009, October). Visual opinion analysis of customer feedback data. In Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on (pp. 187-194). IEEE.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Perner, L. (2005). The psychology of consumers: Consumer behavior and marketing. Journal of Child Pyschology, 32(10), 1-29.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71, 2001.

Posner, M. (2012). Very basic strategies for interpreting results from the Topic Modeling Tool. Retrieved May 29, 2015, from http:/lmiriamposner.comlhlog/very-basic-strategies-for-interpreting- resuIts-from-the-topic-modeling-tool/

Robertson, T. S. (1967). The process of innovation and the diffusion of innovation. The Journal of Marketing, 14-19.

Ritterman, J., Osborne, M., & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. Paper presented at the 1st international workshop on mining social media.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. ACM Transactions on Information Systems (TOIS), 28(1), 4.

Romero, D. M., Meeder, B., & Kleinberg, J. (2011, March). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In Proceedings of the 20th international conference on World wide web (pp. 695-704). ACM.

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 dutch senate election results with Twitter. Paper presented at the Proceedings of the Workshop on Semantic Analysis in social media.

Feldman, R. and Sanger, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, Cambridge, U.K., 2007.

Schumpeter, J. A. 1934. The Theory of Economic Development. An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle. Cambridge, Massachusetts. Harvard University Press.

Scott, J. (2012). Social network analysis. Sage.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PloS one, 6(5), e19467.

Surowiecki, J. (2004), The Wisdom of Crowds, Doubleday, New York.

Steinwart, I. and Christmann, A. *Support Vector Machines*. John Wiley & Sons, Inc., New York, 2008.

Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology, 29(1):24–54, 2010.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12) (2010) 2544-2558

Thomas, J.J. and Cook, K.A. A visual analytics agenda. *IEEE Computer Graphics and Applications 26*, 1 (Jan./ Feb. 2006), 10–13.

Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics* (Philadelphia). ACL, Stroudsburg, PA, 2002, 417–424.

Tidd, Joe & Bessant, John & Pavitt, Keith. (1998). Managing Innovation: Integrating Technological, Market and Organizational Change. West Sussex, England: John Wiley & Sons Ltd.

Trott, P. (2008). Innovation management and new product development. Pearson education.

Tumasjan, T. Sprenger, P. G. Sandner, and I. M. Welpe, “Predict- ing elections with Twitter: What 140 characters reveal about political sentiment,” in Proc. of 4th ICWSM. AAAI Press, 2010, pp. 178–185.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time Twitter sentiment analysis of 2012 u.s. presidential election cycle. In ACL System Demonstrations, pages 115–120, 2012.

D. Watson and L. Clark. Development and validation of brief measures of positive and negative affect: the panas scales. Journal of Personality and Social Psychology, 54(1):1063–1070, 1985.

Wolfers, J. and Zitzewitz, E. (2004), “Prediction markets”, Journal of Economic Perspectives, Vol. 18 No. 2, pp. 107-126.

Yin, Z., Cao, L., Gu, Q., and Han, J. Latent community topic analysis: Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology 3*, 4 (Sept. 2012), 1–21.

Yu, S., & Kak, S. (2012). A survey of prediction using social media. arXiv preprint arXiv: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, 55-62.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). social media analytics and intelligence. Intelligent Systems, IEEE, 25(6), 13-16.