REPUBLIC OF TURKEY
FIRAT UNIVERSITY
THE INSTITUTE OF NATURAL AND APPLIED SCIENCES

ANALYSIS OF OPINION LEADERS USING TEXT MINING TECHNIQUES ON SOCIAL MEDIA

MASTER THESIS

Kaloma Usman Majikumna
(141137105)

Department: Software Engineering
Supervisor: Asst. Prof. Dr. Mustafa Ulaş

June-2016

REPUBLIC OF TURKEY
FIRAT UNIVERSITY
THE INSTITUTE OF NATURAL AND APPLIED SCIENCES

ANALYSIS OF OPINION LEADERS USING TEXT MINING TECIINIQUES ON SOCIAL MEDIA

MASTER THESIS

Kaloma Usman Majikumna
(141137105)

Date Submitted to Institute: May 9, 2016
Thesis Defense Date: June 2, 2016

Thesis Supervisor: Asst. Prof. Dr. Mustafa Ulaş (F.U.)
Other Jury Members: Prof. Dr. Asaf Varol (F.U.)
Assoc. Prof. Dr. Muhammed Fatih Talu (I.U.)

June-2016

DECLARATION

I declare that this thesis with title “Analysis of Opinion Leaders Using Text Mining Techniques on Social Media” is prepared by myself as a partial fulfillment of the requirements for the degree of Master of Science in Software Engineering.

Kaloma Usman Majikumna

ELAZIĞ-2016

DEDICATION

To the family of Malam Usman Bulama Majikumna with all my love.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to my thesis supervisor Asst. Prof. Dr. Mustafa Ulaş for his advice, guidance, supervision, patience and for all his kind support. I will always remember his support especially for providing me with a special place at his office, a computer and all the necessary materials at my disposal to ensure the success of this research. My thanks to Res. Asst. Osman Altay for assisting in editing the format of this thesis.

I wish to thank Software Engineering Department Chair Prof. Dr. Asaf Varol for making the admission procedure easy and for all his encouragements during the coursework and the research; I would also like to thank him and the other jury members of my thesis. My gratitude also goes to the Turkish people especially those in Elazığ for their hospitality. I would also like to thank all friends and faculty members that have participated in this thesis in one way or the other.

I will always be grateful to our late Governor Mamman Bello Ali who is the first person that initiated and supported my scholarship in the Republic of Turkey. I am grateful to my friend Engineer Serkan Özkan for his kindness during my studies in the Republic of Turkey. Finally, I would like to thank my family for their understanding, patience, encouragement and all kind of support. This research wouldn't be possible without their love, care, prayers and guidance.

Kaloma Usman Majikumna

TABLE OF CONTENTS

Page No

ABSTRACT

Analysis of Opinion Leaders Using Text Mining Techniques on Social Media

The advent of social media technology has witnessed rapid adoption globally. There are many ways in which people engage in using this technology. Some people use this technology for fun, some for commercial purpose, while others use it for education. The social media has become essential tool in the lives of people. Nowadays, people can check what other customers think about a product or service before they buy on the Internet. Opinion can be formed referring to shared feedback on the virtual community which redirect the new opinion of people.

There are few individuals who can influence buying decisions, ideas or even political ambitions of others through the comments they made on the social media. Those individuals are usually referred to as opinion leaders or influential leaders. It is quite possible to reach out and communicate with a large number of people on the social media through the opinion leaders. The importance of these opinion leaders in marketing made companies start to change their marketing strategies. However, identifying such opinion leaders is not as easy as it may seem. So, the identification and analysis of opinion leaders on social media is a valuable research.

This study identifies and analyzes the opinion leaders on Twitter social media site. The first method used to identify the opinion leaders is the traditional measure of influence called indegree, the second one is the retweets criterion, and finally a measure of influence based on sentiment analysis is proposed. The sentiment analysis method uses Naïve Bayes classifier for the classification of sentiments polarity and it has an accuracy of 59.51%. The analysis based on retweets and sentiments showed that having a high indegree rank (the traditional ranking score) does not necessarily mean a user is an opinion leader.

Keywords: Influential Leaders, Naïve Bayes, Opinion Leaders, Opinion Mining, Text Mining

ÖZET

Metin Madenciliği Teknikleri Kullanılarak Sosyal Medya Verileri ile Kanaat Önderlerinin Analizi

Sosyal medya teknolojisinin hızlı gelişimi ile dünya çapında benimsenmesine tanıklık edilmiştir. Birçok sebepten ötürü insanlar sosyal medya ve beraberinde gelen teknolojileri kullanmaktadır. Bazıları öğrenme için, bazıları eğlence için bazıları ise ticari amaçla bu teknolojiyi kullanmaktadır. Sosyal medya insanların yaşamlarını yönlendiren önemli bir araç haline gelmiştir. İnsanlar internet üzerinde herhangi bir ürünü satın almadan önce sosyal medya gibi birçok kaynaktan araştırma yapmaktadır. Bu ise insanların paylaştığı fikirler ile fikir bulutlarını oluşturmakta ve insanların yeni oluşturdukları fikirleri yönlendirmektedir.

Sosyal medyada fikir liderlerinin yaptığı yorumlar aracılığıyla, başkalarının satın alma kararlarını, fikirlerini, hatta siyasi emellerini etkileyebilmesi muhtemeldir. Bu fikir liderleri genellikle kanaat önderleri (fikir liderleri) ya da etkili liderler olarak adlandırılırlar. Kanaat önderleri aracılığıyla sosyal medyada birçok insanla iletişim kurmak ve yönlendirmek mümkündür. Pazarlama şirketleri bu kanaat önderlerinin önemini fark ettikten sonra stratejilerini değiştirmeye başlamıştır. Ancak bu fikir liderlerinin tespiti göründüğü kadar kolay değildir. Bu yüzden sosyal medyada liderlerin tespit ve analiz edilmesi hakkındaki çalışmalar değerlidir.

Bu tez çalışmasında Twitter sosyal medya sitesinde kanaat önderlerinin tespiti ve analizi yapılmaya çalışılmıştır. Fikir liderlerini belirlemek için kullanılan ilk yöntem, kullanıcının takipçi sayısıdır. İkinci kriter olarak ise retweets kullanılmıştır. Son olarak duygu analiz yöntemiyle fikir liderleri tespitine çalışılmıştır. Duygu analizinde sınıflandırma için kullanılan Naive Bayes algoritması ile %59.51 doğruluk elde edilmiştir. Yapılan araştırmalar sonucunda, retweets ve duygu analizine göre bir kullanıcının çok takipçisi olması o kullanıcının çok etkili olduğu anlamına gelmediği görülmektedir.

Anahtar Kelimeler: Fikir madenciliği, Fikir Önderleri, Kaanat Önderleri, Metin Madenciliği, Naive Bayes

TABLES LIST

Page No

SYMBOLS AND ABBREVIATIONS

TM : Text Mining

DM : Data Mining

SVM : Support Vector Machine

ML : Machine Learning

K-NN : K-Nearest Neighbors

AI : Artificial Intelligence

WOM : Words of Mouth

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

NLP : Natural Language Processing

F.U. : FIRAT UNIVERSITY

I.U. : INONU UNIVERSITY

FIGURES LIST

Page No

1. INTRODUCTION

Social media technology has been improving rapidly in recent years and people's access to social media via the Internet is increasing each and every single day. The social media technology and the Internet technology are developing in parallel. The availability of Internet makes people change their way of life from physical interactions to interactions via the virtual community. Nowadays, due to the availability of social media sites such as Twitter, Facebook, and Instagram etc. it is quite common and easy to interact and communicate with a diverse number of people within and across national boundaries.

Before the advent of social media technology, people used to have a limited number of friends or relatives to communicate with and seek for advice from them. But the virtual community makes it possible for individuals to interact and make beneficial connections with peoples who have similar interest. It is almost easy for all groups of people to find somebody that has a common interest in their day-to-day activities with the help of the virtual community or the social media. For example, social media is now a means for businesses to buy and sell things, and it is a platform for students and academics to advance their knowledge by tracking, communicating and collaborating with their peers in other parts of the world. Such communications were not easy, if not impossible at all, without the social media sites.

There is no doubt that physical interactions among individuals are very important, but there are many benefits of interaction beyond the physical domain. Traditionally, people rely on their friends, family, and neighbors for advice concerning everything they lack sufficient knowledge to decide or choose the best of it. Edward and Berry (2003) reported that 83% of people refer to their friends, family or someone knowledgeable at a restaurant before they choose where to eat what over the traditional restaurant advertisement, 71% of people also refer to their family, friends or an expert before visiting a place or buying a prescribed drug, and 61% of people do the same before they watch a movie [1].

Generally speaking, influencing others by means of communication is what has been broadly referred to as the words of mouth (WOM) by researchers. Johan (1967) stated that WOM is basically a process of diffusing or spreading the word about a service or a product from person-to-person i.e. between a communicator and a receiver and the receiver perceives the information as noncommercial [2]. One of the pioneering research on the WOM diffusion reported by Katz and Lazersfeld (1955) describes influential individuals as "opinion leaders", the opinion leaders are the enlighten small number of people in a community that intercept information from mass media, interpret the information, and diffuse the information they received to the personal networks that they themselves belong [3]. The term opinion leaders are generally accepted as those individuals whose opinions concerning several issues or things are widely accepted and utilized by the people in their respective community [4].

The above explanations lead us to what is currently evolving in today's world; as explained above, Internet usage is essential in people's lives; they can buy and sell things or services via the Internet. Likewise, people can check what others think about a product or service from other consumers on the Internet. For example, if a person wants to buy a product from Amazon.com or Alibaba.com, the first thing that person or customer can usually do is to check the reviews for that product and see what others think about the product, however traditionally a consumer will only ask within his community or close contact and learn if someone knows more about the product so that he or she can help him decide whether to buy or not. The buyer or consumer usually buys the product or service if his or her friend or family member has suggested that product. Many advertising companies agree with the idea of influential individuals on the social media, this leads to shifting their focus to advertising via the influential people.

Influential individuals are not limited to influencing audience for buying decisions only; they also can influence mass audience for election campaign i.e. during electioneering and many other issues. Today, there are many such influential individuals or opinion leaders on the social media; they can easily influence a large number of people on social media by writing comments. Agarwal, Liu, et al (2008), stated that identification of such influential individuals can help in forging political agenda, developing innovative business strategies, discussing social issues and it can lead to many interesting applications [5]. For example,

influentials often affect buying decisions of their followers, so, companies can rely on them to some extent in order to advertise and influence their consumers.

Drezner and Farrell (2004) proposed that the influentials could also sway public opinions on reactions to government policies, elections, and political campaigns [6]. It is a well-known fact that public opinion is very important for each and every organization and government body, thus, even the government could refer to such opinion leaders in order to reach out to it citizens on arising issues.

The discussion of opinion leaders on social media would not be sufficient without talking about text mining. Text Mining (TM) was broadly defined by Feldman, Sanger (2007) as a process by which a user interacts with documents' collections by using analysis tools over time. In a manner similar to data mining, TM aims to extract beneficial information from several data sources by exploration and identification of interesting patterns. In the TM perspective, data sources are collections of documents, and interesting patterns are from unstructured text data, but not in formalized database records [7]. Expressing opinions and sentiments in written form on social media are very common among individuals, organizations, businesses, consumers, and celebrities. There are many applications or importance of TM in our daily lives. TM applications can help businesses and corporations to get feedback from their targeted consumers in order to know exactly how to improve the quality of their products or services [8]. A previous method of getting users and customers generated feedback is a questionnaire, there is no doubt that it is easier to collect and analyze users and customers generated feedback on social media than using questionnaire. TM can help in retrieving unbiased customers reviews or citizens' reviews through the social media. Sentiment analysis of people's comments on social media sites such as Twitter or Facebook can easily clarify if consumers are satisfied with products or not.

Text mining on social media is a relatively new research field; it adopts most of its techniques from artificial intelligence, machine learning, and data mining. Famous machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes, Artificial Neural Networks, and K-Nearest Neighbor (K-NN) etc. have been used for classification and clustering purposes. In this study, Naïve Bayes classification algorithm is used to classify and analyze the sentiment polarities of Twitter users.

In the literature, there are many works on opinion leadership detection and analysis on social media. Some of these works propose methods using network analysis while others using text mining methods similar to this research, the details are given in the background and related works section of this thesis.

This research aims to analyze opinion leaders on social media using text mining techniques. Social media is such a large term, to be precise Twitter microblogging site will be used in this study. The proposed solution does not depend on Twitter data, it could also be used for Facebook and many other social network sites, and it is just that the data need to be normalized to be compatible with the proposed solution. In addition, at the end of this thesis, text mining application will be developed in order to fully understand the general concept of opinion leader and sentiment analysis based on real life data that will be retrieved from the social media.

The Introduction chapter of this study introduces the main idea behind this research; it gives information about the opinion leader or influential leader and its importance. It also introduces text mining usage for better understanding of the research. In Background and Related Work chapter, the background field of the research is given; the chapter introduces Artificial Intelligence (AI), Data Mining (DM), Machine Learning (ML), and elaborates Text Mining (TM) and its applications area. The chapter also presents the opinion leader related work and gives some analysis. Application for Finding Opinion Leaders on Twitter chapter gives the general structure of the methodology and the approaches which are used in this research. In Results and Discussion chapter, the general analysis and the discussions about the findings is presented. In Conclusion chapter, the overall conclusion of the thesis and possible future research is presented.

2. BACKGROUND AND RELATED WORK

This section aims to give a general background work and a general literature review related to the opinion leaders. There are mainly two subsections i.e. the background section and the related work section.

2.1. Background

Analysis of opinion leader using text mining on social media is a relatively new research field, text mining research field itself is a new research field. So, there are some background terms and topics that are very important to be explained. It is essential to comprehend the basic concept of Artificial Intelligence (AI), Machine Learning (ML) and that of Data Mining (DM) for a better understanding of Text Mining (TM). As the discussion goes on, it will be seen that there are many things in common between AI, ML, DM and TM, though TM is very specific research field that mainly concentrated on text data.

2.1.1. Artificial Intelligence

Artificial intelligence is one of the most important research fields in computer science and engineering. Many researchers proposed several definitions of AI in the literature, among these definitions:

- AI has been defined by Charniak and McDermott (1985) as “The study of mental faculties through the use of computational models” [9].
- Another definition by Schalkoff (1990) described AI as "A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes" [10].
- Winston (1992) defined AI as "The study of the computations that make it possible to perceive, reason, and act" [11].
- Luger and Stubblefield (1993) stated that AI is “The branch of computer science that is concerned with the automation of intelligent behavior" [12].

There are many applications of AI that have been described by researchers; some of them are described by (John McCarthy, 2007*)* as follows: Game Playing,

Speech Recognition, Understanding Natural Language, Computer Vision, Expert Systems and Heuristic Classification [13]. So, AI is directly related to TM through natural language processing.

2.1.2. Machine Learning

Machine learning is one of the leading research fields in computer science and engineering. Machine learning and artificial intelligence share many things in common. In fact, machine learning can be considered as a subtopic in AI. Some of the definitions of machine learning reported by researchers are as follows:

- In particular, Kevin Murphy defined machine learning "as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)" [14].
- Machine learning as defined by Arthur Samuel (1959) "is the field of study that gives computers the ability to learn without being explicitly programmed" [15].
- Next definition of machine learning was stated as "Machine learning is programming computers to optimize a performance criterion using example data or past experience" [16].

Some of the applications of machine learning include the following: Learning Associations, Classification, Regression, Unsupervised Learning and Reinforcement Learning etc. [16].

2.1.3. Data Mining

Generally speaking there are many proposed definitions for data mining and knowledge discovery, data mining (sometimes called data or knowledge discovery) as reported by Hitesh Gupta (2011) "is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both" [17].

The application field of data mining as reported by Fayyad, Piatetsky-Shapiro and Smyth (1996) includes the following: Marketing, Investment, Fraud Detection, Manufacturing, Telecommunication, and Data Cleaning [18]. Based on the above brief

discussion on Artificial Intelligence, Machine Learning, and Data Mining, it can be inferred that Text Mining comes about from the field of Artificial Intelligence, Machine Learning, and Data Mining. Figure 2.1 below shows the picture of Text Mining overlapping fields.

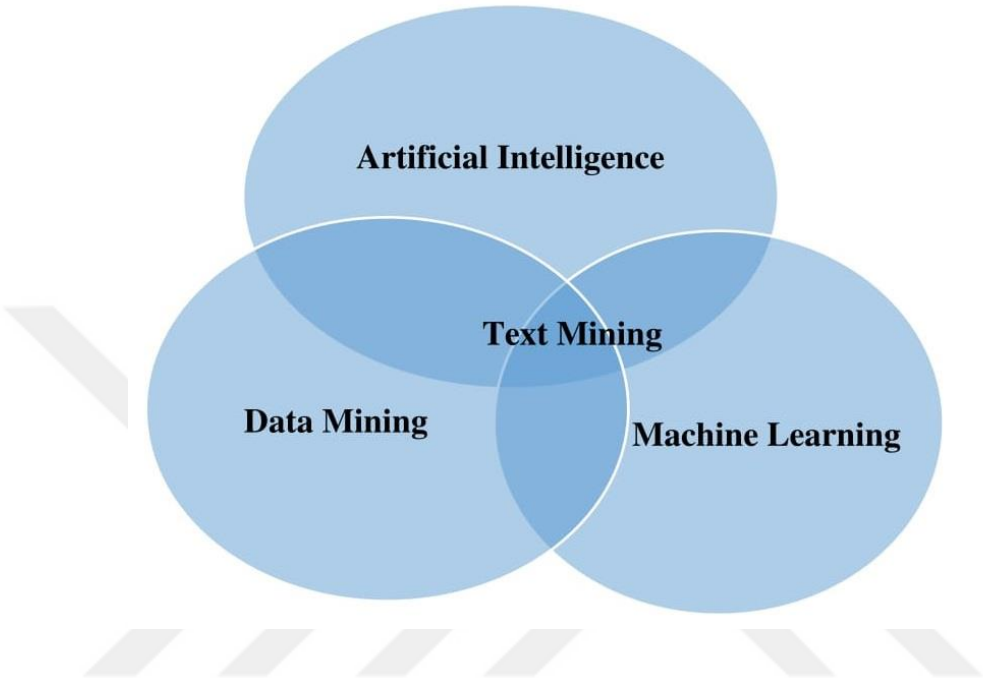

Figure 2.1. Text mining overlapping fields

As explained in the introduction section of this study, there are many fields from which the algorithms used in text mining came from. For example, the Naïve Bayes classification is originally a machine algorithm, but it is used in text mining field for sentiment classification purpose.

2.1.5. Twitter

Twitter is a social network site with about 310 million active users monthly [19]. In Twitter, users can follow others and they can be followed as well. A Twitter user that follows less number of people and has a large number of followers is usually considered to be famous or influential. There are quite important Twitter terms that will be mentioned frequently in this study. So, there is a need to understand those terms clearly. The terms are Tweets, Replies, Mentions, and Retweets.

These Twitter terms, as describe by the Twitter.com [19], are as follows:

- Tweet is a 140 character long text posted on the twitter microblogging site. A user has to express his or her thought with 140 characters only.
- Reply is a Tweet that a user receives from another user in response to his or her specific Tweet, the Reply starts with @username.
- Mention is any Tweet that contains a user's Twitter username starting with @; just like this: Hello @kalomausman, what's up? Mentions are of two types: The first one places the username in the beginning of the tweet and the second one places the username in the middle of the text.
- Retweet refers to reposting of another user's tweet(s). Retweet usually shows agreement by sharing the exact Tweets someone else posted.

In this study, the number of followers (also called indegree) of Twitter users, Retweets influences, and sentiments of Replies and Mentions are considered.

2.2. Related Work

Research interest in detection, recognition, and analysis of opinion leaders on social media has increased tremendously due to the advancement of Web 2.0 or the Internet. Recently, there are a lot of research works on opinion leaders by researchers globally. Some of the researchers concentrate on providing new algorithms based on mathematical formulas and optimization methods for the detection and recognition of opinion leaders on the social network. Other researchers consider text mining and network analysis techniques for the opinion leadership analysis. However, the detailed methods, techniques, approaches and data used differ from one another.

One of the famous and detailed researches on identification and detection of influential bloggers or opinion leaders on social media in the literature was conducted by Nitin, Huan, et al. (2008), they presented a method for identifying influential bloggers in the virtual community or blogs in their research. The authors stated that not all active bloggers are necessarily influential; in other words inactive bloggers may be influential bloggers. A blogger is considered influential if one or more of his blog posts are influential. Four features are used in order to determine influential bloggers' post or a post that is influential. Firstly, the recognition; an influential post gets much recognition from bloggers. Many bloggers made in-links to an influential blog post from their blogs. Secondly, the generation of activity (activity generation); an influential post generates many activities such as discussions and arguments that are posted as comments on the post.

Thirdly, the novelty property; influential blog post usually has little or no external out-links to some blog posts. Finally, the eloquence; an influential blog post is usually long, well articulate with lots of information [5].

A model of an influential graph was proposed by the mentioned authors of the previous paragraph. The model is according to the mentioned properties above. TUAW (unofficial apple weblog) was used in the process of model comparison. The results of active bloggers' list on TUAW and the results of the proposed model were compared. It was shown that an active blogger is sometimes an influential blogger and sometimes not, the same goes for influential blogger. The results of Digg social media platform was also used for the evaluation. Blog posts are given higher points if they have many likes. According to the results, it is concluded that almost all influential blog posts have low out-links, more comments and are longer lengths. The results also show that weights of the properties used in the proposed model are the important key points. So, if the weights are altered, the results change immediately. The authors expect that this model may serve as a starting point in identifying influential bloggers and the parameters used should be considered for future research [5].

Bodendorf and Kaiser (2010) detected trends and opinion leaders on social media using network analysis and text mining techniques. Four steps were used in their approach. The first step is the opinion recognition using text mining techniques. They extracted the features from the text and then applied the text mining algorithms for the classification based on the polarity of the data. The authors used ternary classification with positive, negative and neutral classes respectively. The method used for the classification was based on supervised learning. The second step is the identification of the relationships among the entities. To achieve that, relationship extraction techniques and co-reference resolution were used. Co-reference resolution means recognizing whether two distinct words in a single text refer to the same object or not. After the two steps mention above, social network analysis was applied and the concept of "centrality" was taken into consideration in order to detect the opinion leaders. Betweenness, closeness and degree centralities were considered as the centrality metrics during the process. Afterward, the opinion leaders were detected, followed by trend detection. For the analysis of the opinion leaders, the properties of the network including closeness centrality, density, and Randic connectivity (the level of network branching) were considered. The lower the Randic connectivity, the higher the

density, the more the forum users are connected with one another and communicate without intermedium object. In addition, centralization showed whether one trend arises from an opinion leader or from a different one. During the experimental work, threads extracted from pcfreunde.de and slashfm.de forums were used [20].

Wu and Ren (2011) conducted a research on sentiment influence analysis on Twitter using twits, retweets, mentions, and replies. The authors collected around 3200 tweets from 1000 twitter users and applied sentiment analysis method. Unsupervised lexicon binary classification was applied in order to classify the text. The tweets were classified as positive or negative based on the number of positive or negative words that appear in the text. Thereupon, the researchers defined an influence model of probability; the model was defined based on three entities or characteristics: Mention, Reply, and Retweet. They combined the model with the results of the sentiment polarity found in the previous step. In order to determine the sentiment influence, they computed negative user to user influence probability by using only negative mentions, replies and retweets over the combined negative tweets published by the users, and they did the same for the positive ones [21]. Their work resembles our proposed method in terms of using text mining techniques i.e. sentiment analysis in this case and using text mining techniques for opinion leaders detection. However, there are differences in the methodology. Our approach ranks users based on indegree, retweets, and then finally based on sentiments responses of other Twitter users.

Duan, Zeng, and Luo (2014) proposed a system in order to identify opinion leaders based on user clustering techniques and sentiment analysis. The authors' proposed a model that adopts two main stages in the identification of the opinion leaders. They first generated the candidate opinion leaders and the final selection. Other processes needed to include preprocessing the data and sentiment extraction. Two datasets i.e. stock trading information and stock message board were used in the experiment. According to their framework, the mechanism of the identification of the opinion leaders in the stock message boards is as follows:

- Preprocessing was applied; this included the omission of stop words, URL, and message filtering.

- The authors clustered the users into several clusters based on the below rules; after clustering, candidate opinion leaders were generated.

- After preprocessing and clustering sentiment were extracted. Opinions about price trend on the posts from the message boards were extracted, and lastly, a new dataset that contains users' sentiment, date, and the stock were generated for next step processing.

- A method for selection based on correlation was applied in order to evaluate the correctness of all users' opinions on the stock price movement trends.

The selection process has sentiment dataset and stock trade data as the input, after that correlation coefficient was calculated for all the candidate opinion leaders. Consequently, the final opinion leaders were selected based on the correlation coefficient. This correctness reflects the users' ability to analyze the stock market. Initially, the authors consider four rules for the selection of the candidate opinion leaders. The rules include:

- The message posted by the opinion leader has to be little long.

- The average number of replies to the opinion leader's post has to be relatively high.

- Opinion leader's reply to other users has to be moderate.

- Finally, the total post written by the opinion leader should be moderate.

Considering the general authors' method here [22], there are few things that are missing and may affect the performance of the proposed method. Such as, rule 3 and 4 may hold without any issue, however, rule 1 may be modified and make the system consider short or long post in selecting candidate opinion leader; another one is in rule 2, sentiment analysis can be applied in rule 2 such that opinion leader's post should have more supporters that agree with his post based on the reply he or she receives.

Another work by Joshi, Finin, Java, et al. propound a system that analyzes social media applications and systems in order to be able to find opinions on specific interest topics, recognize spam blogs and to detect opinion leaders using trust relationships. Analysis of Blogosphere is performed in the process of conducting their research for it richer network structure. Blogosphere's richer network structure comes from several relationship types

that exist in between bloggers and also between blogs. The example of these relationships can be co-occurrence, outdegree, indegree, and reference relations. An influence on Blogosphere is shaped by taking into consideration the fact that most of the time it is the function of a specific topic. The authors also mentioned that influence is measured correctly by considering opinions and the sentiment. But the obstacles that prevent or reduce the high accuracy during identification of these influential users are spam comment, spam blogs, non-content blog-rolls, advertisements, and link-rolls. To solve these issues or obstacles, the researchers proposed kind of techniques using relational and local features of blogs. As a result, the spams were detected; the authors then used the idea of opinions polarization. The sentiment determines the link polarity of the text data that surrounds the whole links that connect the blogs. The authors used sample dataset of 1490 blogs that had been labeled as republican and democrat for each blog. After the experiments, their findings revealed that classification using polar links produced better results in comparison to plain link structure. The authors also revealed that the rapid increase in the complexity and massive usage of the social media bring new challenges and opportunities for research [23].

A research on finding influential leaders from opinion networks was conducted by Zhou, Zeng, and Zhang (2009). The authors used sentiment information in a text instead of widely used centrality properties of graphs in opinion networks. The proposed algorithm is similar to PageRank algorithm. In fact, the researchers modified PageRank algorithm and called it OpinionRank. They used OpinionRank in finding opinion leaders. As an initial stage, the authors extracted sentiment information and strength with the help of machine learning techniques. After that, opinion network was defined based on implicit opinion orientation, all the graph nodes participate in communication and some social activities, and they all form their opinions on one another as time goes. The proposed method uses sentiment information by consolidating opinion scores and integrating the famous PageRank algorithm according to these scores. The nodes in the opinion network were ranked based on their influences. Data from epinions.com were used in the process. One interesting issue is that the performance of PageRank and OpinionRank is almost the same even though the authors modifid the latter. At the end, Activity based method appear to be the best among others [24].

Next proposed method for influence detection in blogs is based on Natural Language Processing (NLP) by Hui and Gregory. The authors tried to quantify influence and sentiment by using the combination of sentiment analysis and NLP in blog spaces. The proposed method investigated blogs based on topics. The authors in this work used NLP to classify blogs into several topics. They studied the comment sections of these topic-specific blogs based on their sentiments, sentiment analysis methods were used in the process. The researchers classified the sentiments, and then the influences were calculated with respect to four specific criteria: Number of followers, comments, relevancy, and followers. The last criterion shows that if a potential influential blogger has a follower who is influential on a similar topic, there is a greater chance that this potential influential blogger is also an influential blogger. So, they proposed the algorithm with the mentioned four criteria. The algorithm is similar to Google PageRank algorithm. There is an important connection between the sentiment of the post and the sentiments of the comments, but the authors chose to ignore it in order to simplify the proposed method [25]. So, the mentioned connection between post sentiment and comments sentiments could be explored in future research.

Other researchers showed how to analyze social media using social network analysis and using text mining techniques. The proposed work by Merwe and Heerden (2009) revealed how to find opinion leaders by using social network theory and opinion leadership together. As in most of the work in the literature, this research showed that people that are at the network center are the opinion leaders. It is also found out that general leadership is strongly related to domain-specific (topic specific) leadership unlike in other studies. Furthermore, it is reported that a strong correlation exists between opinion leadership mentioned by others and self-reported (self-claim) opinion leadership. The researchers administered a survey questionnaire during the research for data collection from five groups of college students; each group consists of 25 students approximately. The first criterion used to determine if an individual is opinion leader is that students were asked to give a point for their classmates that they can rely on for some suggestions, the students which received a high number of nominations were considered as the opinion leaders. The second criterion of determining the opinion leadership is the self-reported opinion leaders; self-reported opinion leadership was calculated based on the feedback of the respondents (self-claim opinion leaders). The authors highlighted two issues in their work [26]:

- The data used is very small, there is a need to use a huge amount of data so that more reliable results can be achieved [26].

- The participants in the research were all students, and students are sometimes considered to be less reliable for gathering such information; there is a need to conduct the research with experienced people in the market or industry [26].

In this study there is a limitation of the work being for only one specific domain, so, the authors suggested that future researcher could study the link between non-domain specific and domain specific opinion leadership. Our proposed method for opinion leadership is not domain specific, its general opinion leader detection based on sentiment analysis.

According to a detailed and comprehensive study made by Kardara, Papadakis, et al. (2012) on influence pattern on social media, there are two theories of influence diffusion. The first one is the minority theory; it states that people that are very popular yet unrelated influence many other people. The second one is that people are affected and influence by their peers, as a result, all are influential. The researchers' goal was studying dynamics of influence in specific topic context on Twitter. They chose to work on Twitter due it flexibility and easy access to user data for such academic research. They studied four different influences: tweet, retweet, indegree and mention influences. Tweet influence refers to the total number of posts a user posts on a particular topic. Retweet influence refers to the total number of tweets of a user that are retweeted by others. Mention influence is the number of mentions a user receives from others when they refer to him/her. Indegree influence is the total number of followers a particular user has in a community. The authors used these four metrics to form different groups. They presented several influence patterns from the four metrics, those patterns help in evaluating influential users or candidate opinion leaders. Among the patterns is Cross-Criterion Overlap Pattern. They tried to assess if different metrics classify identical users as opinion leaders (or influential users) based on the same topics. The famous Jaccard similarity metric is used to gather and get the overlaps in the groups. The second pattern used in this work is Sentiment Pattern. They propound that sentiment of the whole community should be determined by the potential influencers. To achieve that, Pearson correlation coefficient and polarity ratio is used. The third pattern used is Content Volume Patterns which investigates community's

total activities within its center and core groups. The fourth is the Structure Patterns which is used to evaluate the degree of a common background of core users and their opinions. Last two patterns considered in this work are Cross-Community and Intra Core Reference Influence Patterns. Intra-Core Reference Patterns evaluate whether an overlap is found between core groups of two distinct topic communities on same influence criterion. Cross Community Patterns check whether core users are familiar with one another and interact deeply in social matters in respect of the connections they have on the graph network. At the end of their study, the researchers described opinion leaders as users that produce original and factual content that originate from themselves. Opinion leaders also avoid participating in other discussions or using other users' opinion, even though other users retweeted and refer to them often in their discussions. For future steps, they planned to develop a method that will automatically recognize community users that pose the highest influence on others. They also planned to find a way in which the changes in the core groups will be updated in some intervals [27].

The idea of PageRank algorithm is used to propose TwitterRank algorithm by Weng, Lim, Jiang, and He (2010), the authors proposed the algorithm using link structure and topic similarity on Twitter. The PageRank algorithm did not take into consideration both the link structure and topical differences, so this is one of the contributions of their work. Homophily exists among Twitter users as shown by the researchers, Homophily in Twitter means an individual follows someone because he likes the topic(s) that he or she is sharing on Twitter, the second user follows back if he or she thinks that the first person is sharing similar posts. They used 1000 Singaporean based Twitter users for the research. They evaluated the followers and friends (those followed by the 1000 users) of these Twitter users. The tweets of these friends and followers were collected. Topics were extracted for each Twitter user, differences in topics between each and every Twitter users were found and evaluate. Social graph that has "following" relationships among all the users were constructed in the modified PageRank, surfer visits each tweeter user randomly with a probability, and it is a topic based random visit, by following the best edge. The authors claim that they were the first to introduce the concept of homophily on twitter [28].

Another research on influence detection on Twitter reported by Leavitt, Burchard, Fisher, and Gilbert revealed some interesting ideas even though they used a small quantity of test data for their algorithm. The authors evaluated follower to followee ratio,

conversation, and content related feedback. They also defined conversation related responses as the number of mentions and replies and content-related feedback or response as retweets. They tried to find influential Twitter users by computing the responses generated by the users' original posts. They also claimed that follower to followee does not give an accurate measure of influential Twitter users. It is shown that among the signs of influencing an individual are the actions generated an i.e. quantity of reply, retweet, the attribution, and mention received. Their work concludes that celebrities are better at initiating conversations among users while news media sites are better at spreading news content to users [29].

A study on finding influential Twitter users was conducted by Cha, Haddaddi, Benevenuto and Gummadi (2010). This study is similar to the latter research in the previous paragraph above. The authors considered the whole types of relationships on Twitter. They proposed three types of influence called mention, retweet, and indegree influences. They examined the process of spreading news and popular information by these influential users. The authors concluded with the following statements:

- Famous users with high indegree may not necessarily be influential in terms of mentions or retweets.
- Twitter users who get retweeted often also get mentioned often, and vice-versa.
- Most of the influential users often hold a significant degree of influence on certain topics.
- Finally, users are not becoming influential accidentally; users only become influential through consistent personal involvement with their audience [30].

These authors did not consider the sentiments of the mentions, our study further elaborates and considers the sentiments of the mentions for ranking the influentials.

The study presented in [31] is slightly similar to this research is reported by Mustafaraj and Metaxas (2011); the authors conducted a great job in retrieving and analyzing original and edited tweets. They claimed that retweeting original tweets without any alteration shows complete agreement to that author on the specific tweet. They also showed that edited tweets usually reveal opposing motion to that author or twitter user. In this study, a plan was made to analyze the sentiment of the tweets or comments to find out whether the

sentiment support or oppose the twitter user. The authors used some techniques in order to provide labels for users' comments as follows:

- Getting the whole list of who twitter users they follow.
- Checking all twitter users to see whether they have retweeted verbatim another user's tweet in their tweets history.
- They can infer twitter users' political orientation through specific hashtag usage previously.
- Twitter user agrees with original tweets author if she/he retweeted verbatim and follows the author.
- Twitter users that share political orientation, follow one another and used similar hashtags are believed to agree with one another, and vice versa.
- There are two political parties in their analysis, liberals, and conservatives.

The authors discussed the difficulties in extracting comments from edited tweets and analyzed some of the comments [31].

Another research by Chaudhry (2013) discussed opinion leaders and the role they play in consumer purchase decision. The researchers also found that opinion leaders always influence their friends, families, and associates in buying decision. Their findings suggest that marketers should direct their marketing towards identifying opinion leaders so that the opinion leaders can influence or convince non-opinion leaders to adopt their products or services. For example, Google is one of the leading companies among it is competitors such as Yahoo and MSN. But the expenses of Google is far less compared to Yahoo and MSN because Google uses word-of-mouth as a means of advertising compared to its competitors that use other means of advertising [4]. Based on these findings, one can say that utilizing opinion leaders for advertisement can reduce the cost of company or organization.

Finally, Hung and Yeh (2014) proposed text mining techniques in order to identify and evaluate opinion leaders on virtual community. They used features of expertise, novelty and the richness of information found in posts for the identification of the opinion leaders.

The expertise measure shows that a potential opinion leader has rich knowledge in a specific field and loves to involve in several group activities. Novelty is the ability of opinion leader to produce original and quality information. The richness of information is used for the identification of the opinion leader. The proposed model calculates the weights of the opinion leaders based on the average of expertise, novelty, and richness of information. In the process of the research, the authors used 9460 members in the virtual community and supposed that the top ten members are the opinion leaders based on their model. Finally, they compared this proposed text mining approach with four quantitative based approaches in the literature; the four quantitative approaches are involvement, betweenness centrality, degree centrality, and closeness centrality. They found that the involvement approach is the best among the four quantitative approaches. The text mining approach also outperforms the quantitative approach [32]. This work is similar to this research because both consider text mining approach in the opinion leader detection and analysis, but the details are quite different, our work also elaborates further by analyzing sentiment polarity of the posts contents.

3. TEXT MINING

Text Mining (TM) also refers to as text data mining, is the process of discovering hidden meaning or information that is not known previously from unstructured text [33]. TM research field has become very popular recently because each and every day there is an increased availability of unstructured text data on the Internet. People express their feelings on social media websites such as Facebook, Twitter, and Google+ etc. Comments on social media sites are one of the sources of data for mining and extracting new information [34]. TM researchers proposed several techniques for discovering new information from unstructured text that are not known by anyone previously. TM techniques help in Sentiment Analysis, Clustering, Text Summarization, Opinion Leaders, and Trends Discovery [33].

TM gets lots of its features from Machine Learning and Data Mining; for example, algorithms for clustering, classification and text Summarization like Support Vector Machine (SVM), Neural Networks, Cross Validation and Naïve Bayes are from machine learning [33].

Among the important tasks in TM is sentiment analysis. The task in sentiment analysis is usually to classify a given text as positive, negative or neutral. Besides classifying topics into positive, negative or neutral, another task in TM is classifying a given text into predefined classes; for example, SVM or Naïve Bayes can be used to classify a given document or text into class A or B with the help of prepared training data.

There are many benefits or importance of TM applications in our daily life. TM can help businesses and corporation to get feedback from their target consumers in order to know exactly how to improve the quality of their products or services [8]. A Previous method of getting user or customer feedback is a questionnaire. There is no doubt that it is easier to collect and analyze user or customer feedback on social media than using a questionnaire. Sentiment analysis of people's comment on social media sites such as Twitter or Facebook can easily clarify if consumers are satisfied with products or not.

Expressing opinion and sentiment in a written form on social media are very common among individuals, organizations, businesses, consumers, and celebrities. The manner in

which opinion leaders express their feelings or opinions on a particular topic has a great effect on peoples' life such as decision making in politics, quality of product, reliability of information and positive or negative impression of event etc. [35].

3.1. Text Mining Steps

There are usually five steps to text mining as described by Mathiak and Eckstein (2004); the steps include text gathering, text pre-processing, data analysis, visualization and evaluation [36]. Similar steps are used in this study. Similar approaches to TM steps are reported in [37], [38]. There are usually special tools for text collection such as Twitter API and Zapier; Twitter streaming API is used in this study. Then text pre-processing is applied. Text pre-processing includes stop word removal, HTTP address removal, words conversion to lowercase letters etc. Next step is the analysis of the text; it includes text clustering, classification, and summarization. The resulting information retrieved can be placed in management information system for visualization and finally the knowledge will be extracted. Figure 3.1 below shows the text mining steps.

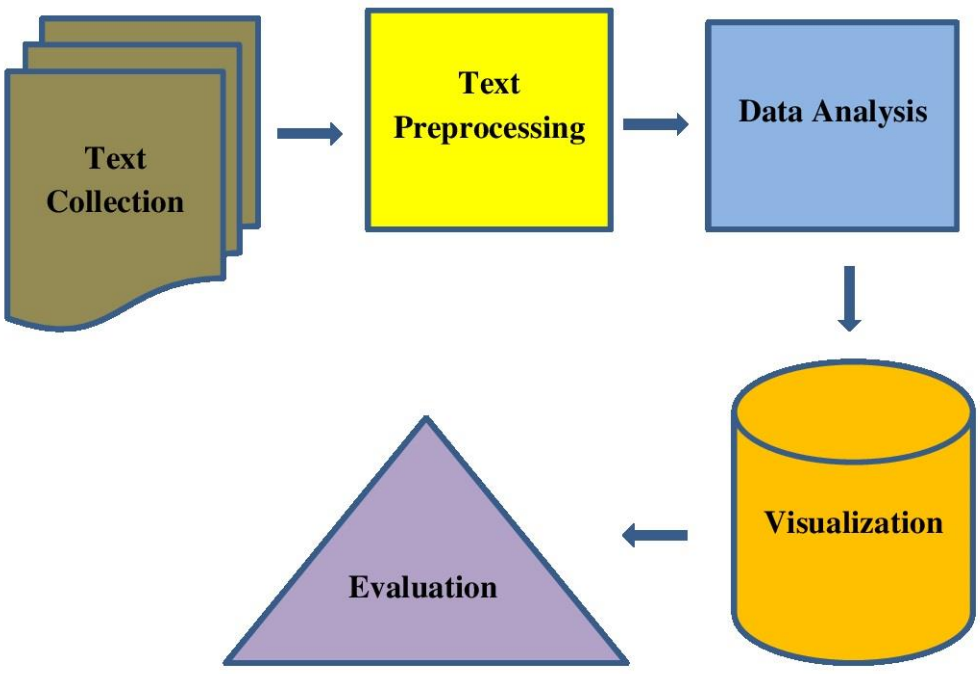

Figure 3.1. Text Mining Steps

The steps used in this thesis start with the text document collection from specific Twitter users to the evaluation step where the knowledge is discovered. The knowledge

discovered here is the amount of positive or negative sentiment a user is perceived by other users on the Twitter virtual community.

3.1.1. Text Data Collection

There are many sources of data for opinion mining; researchers used several social networking sites as a data source for the opinion mining. The source of data for this research is Twitter, Twitter is chosen because there are certain advantages of using it as a source of data for opinion mining. Pak and Paroubek (2010) stated that the following are the benefits of using Twitter as a data source [39]:

- Different people used a microblogging site such as Twitter, so it is a reliable source of people's opinion.
- Twitter contains very huge quantity of text data and it is increasing each and every day.
- Twitter's types of users range from politicians, celebrities, company representatives, and even countries presidents. Thus, it is easy to collect data from different types of opinion leaders and interest groups.
- Twitter is used in many countries though the USA has the largest number of users. It is possible to collect data in many languages; English texts are considered in this research [39].

3.1.2. Pre-Processing

The original text data is retrieved from Twitter users is not suitable for manipulations and other calculations. So, in order to analyze and determine the opinion leadership, some pre-processing techniques are applied as follows:

- Firstly, all the punctuations are eliminated. The eliminated punctuations include: **'_','.','/','*','-','+','(',')','[',']','\','?'** and all other punctuations.
- Secondly, all words are converted to lowercase letters because the written program is case sensitive. For example, "Awesome" and "awesome" are considered to be two distinct words by the program, but these two words convey exactly the same sentiment in a text. So, after the lower case conversion both words become "awesome" and "awesome" which are considered exactly the same word.
- Thirdly, all the text is converted to word list.

The below Table 3.1 tweets samples present the example of the text data before and after the pre-processing techniques.

Table 3.1. Tweets Pre-processing sample

ORIGINAL TWEETS
1. @Cristiano ??????????? so u know how much I?u and @zaynmalik@codylongo ??cause I don't get a chance to see yous I'm sorry I don't get the chance?? 2. I'm not American so this may sound a bit strange but I just realised how much I am going to miss President @BarackObama . #legend 3. For someone who should be such a positive role model...@Cristiano those were poor comments about your teammates
PUNCTUATIONS ELIMINATED
1. Cristiano so u know how much Iu and zaynmalikcodylongo cause I dont get a chance to see yous Im sorry I dont get the chance 2. Im not American so this may sound a bit strange but I just realised how much I am going to miss President BarackObama legend 3. For someone who should be such a positive role modelCristiano those were poor comments about your teammates
TEXTS CONVERTED TO LOWERCASE WORD LIST
1. ['cristiano', 'so', 'u', 'know', 'how', 'much', 'iu', 'and', 'zaynmalikcodylongo', 'cause', 'i', 'dont', 'get', 'a', 'chance', 'to', 'see', 'yous', 'im', 'sorry', 'i', 'dont', 'get', 'the', 'chance'] 2. ['im', 'not', 'american', 'so', 'this', 'may', 'sound', 'a', 'bit', 'strange', 'but', 'i', 'just', 'realised', 'how', 'much', 'i', 'am', 'going', 'to', 'miss', 'president', 'barackobama', 'legend'] 3. ['for', 'someone', 'who', 'should', 'be', 'such', 'a', 'positive', 'role', 'modelcristiano', 'those', 'were', 'poor', 'comments', 'about', 'your', 'teammates']

At the end of the text pre-processing, the texts are now a list of lowercase words that are suitable for all sorts of calculations and manipulations.

3.1.3. Data Analysis

The data analysis is dependent on the preprocessing step discussed previously. Data analysis is a very important step. The actual work is performed here; for example, all the Clustering, Classification etc. take place in this step [36].

3.1.4. Visualization

This part describes the graphical representation sample of the data. Visualization is very important for audience and readers who are going to read about findings of the study. There are three representations samples, one for each of the approaches used for this thesis.

The first one is the indegree distribution of the data, the second is the retweets distribution of the data, and the third one is the sentiment score distribution of the data.

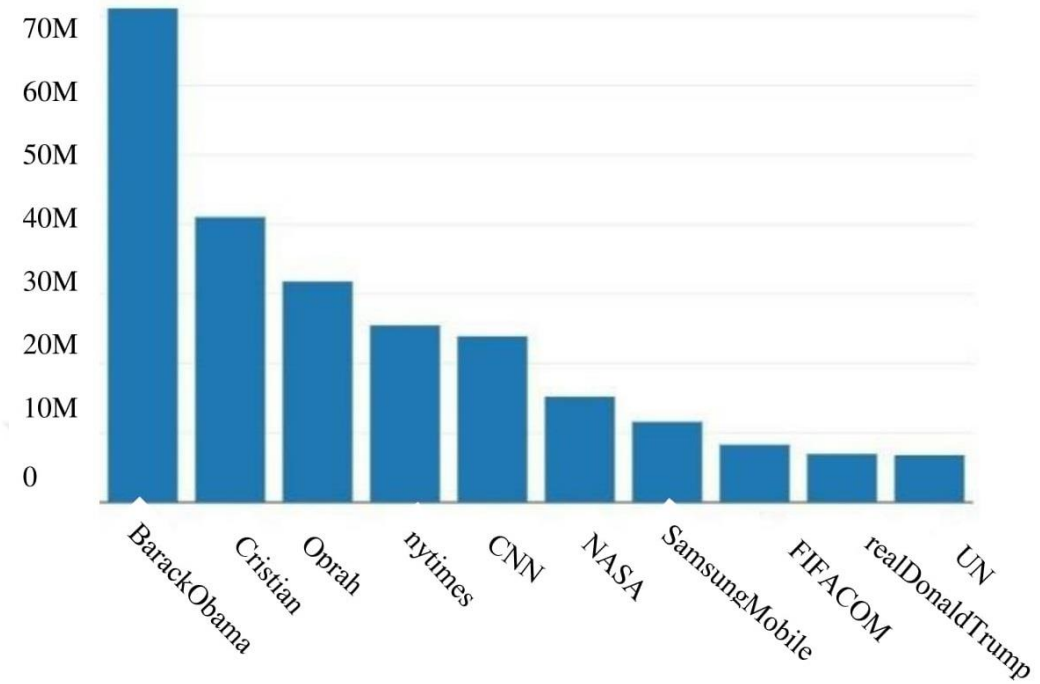

Figure 3.2. Indegree Ranking Representation

The above Figure 3.2 gives the bar chart distribution of the top 10 users based on indegree. This figure shows the user with the highest indegree rank to the user with number 10 highest indegree rank.

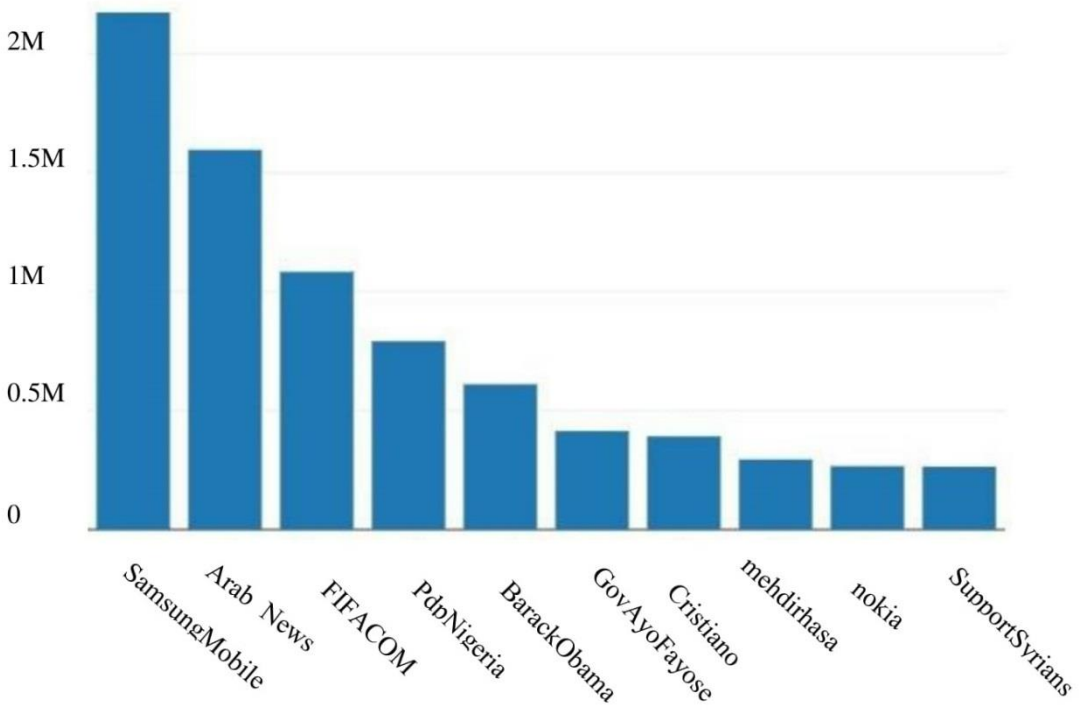

Figure 3.3. Retweets Ranking Representation

Figure 3.3 above shows the bar chart distribution of the top 10 users based on retweets per 100 tweets of each user.

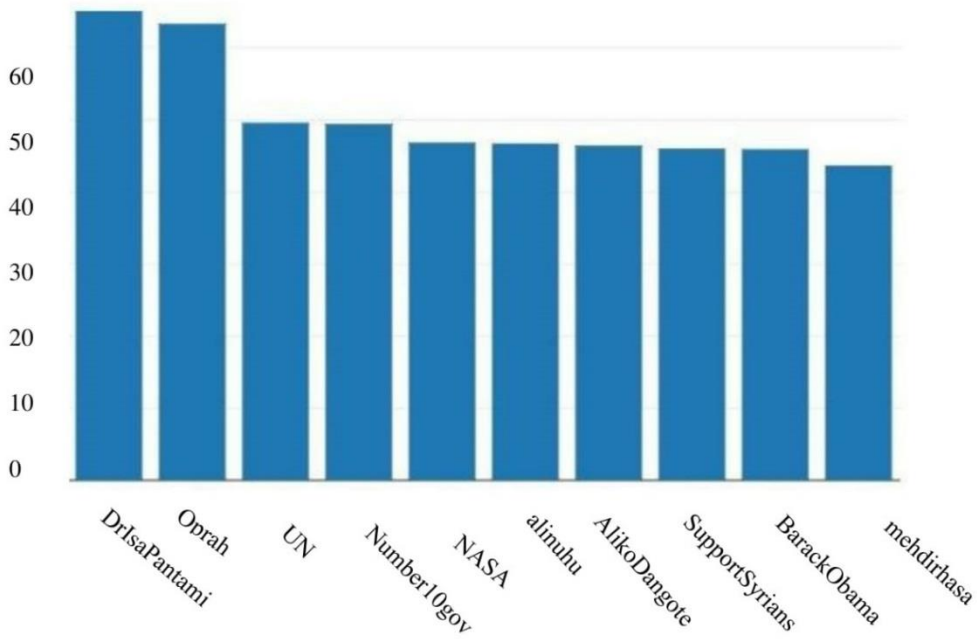

Figure 3.4. Sentiment Ranking Representation

The above Figure 3.4 illustrates the bar chart distribution of the top 10 users based on mentions sentiment score**.**

Based on the three figures (Figure 4, Figure 5, and Figure 6) above, it's quite easy to see how the ranking based on the used methods made clear differences. The analyses of all the methods and approaches are given in the Results and Discussion section of this study. The detailed results of all analyzed 33 users are also given in the Results and Discussion section.

3.1.5. Evaluation

Most of the researches on the opinion leaders or influential leaders found in the literature were domain dependent researches. However, this study is not domain dependent, rather it focuses on detection and the general opinion leadership analysis as already explained in the Introduction chapter of this study. The evaluation is the phase where the opinion leaders discovery and analysis is taking place.

3.2. Application Field of Text Mining

The purpose of applying text mining differs according to the need of the users. As described by Gupta and Lehal (2009) most of the purposes of text mining are as follows [37]:

- Categorization (or Classification)
- Clustering
- Summarization
- Feature Extraction.
- Text-based navigation.
- Search and Retrieval

In addition to the above list proposed by Gupta and Lehal sub-topics were added i.e. Trends and Opinion Leaders. However, discussions were made about Classification, Clustering, Summarization, Regression, Trends and Opinion Leaders.

3.2.1. Classification

The goal of classification is to separate or classify a sentence or document into two or more groups. For example, classifying text as economy news or sports news and

classifying sentiment as negative or positive. The most well-known task in classification is sentiment analysis. The task in sentiment analysis is usually classifying text into Positive Negative or Neutral [40].

Classification is based on supervised learning with predefined training data. The most used algorithm for text classification includes K-nearest neighbor, Neural Networks, Maximum Entropy, Naïve Bayes, and Support Vector Machine (SVM) [40]. According to the findings by Akaichi, Dhouioui and Pérez [34], SVM outperforms other classification methods.

3.2.2. Clustering

Clustering as stated by Kunwar (2013), is the most common and simple unsupervised learning problem and it is defined as "the process of organizing objects into groups whose members are similar in some way" [41]. Clustering plays a similar role as a classification problem, however, clustering is an unsupervised learning problem whereas classification is a supervised learning problem [42]. Classification is suitable for this research.

3.2.3. Regression

Regression is a statistical method for determining a guess or a prediction of unknown variables based on the graphical representation of previous samples. Regression, as defined by Alan, is the analysis or the study of relationship that is usually linear between variables [43].The main formula for linear regression is represented as $y = ax + b$ where y is the dependent variable, and a, b are the independent variables that determine the value of y.

3.2.4. Text Summarization

Text summarization is an important problem nowadays due to the availability of large text data on the Internet. As stated by Radev, Hovy and McKeown, text summarization refers to a text generated from single or multiple text sources, which conveys essential information that is in the original document(s), and which is usually shorter than half of the original text(s) [44].The two different types of text summarization are Single Document and Multiple Documents Summarizations.

3.2.5. Trends

The trend is one of the important tasks in text mining. TM has been used in detecting trends on text data. Nowadays, there is a huge amount of text data on the Internet. For

example, people make comments based on current affairs and emerging several topics of interest on the social media. Trend detection in a collection of text is defined as the detection of the evolving or emerging topic in a text over a period of time [45]. Definition of a trend in the context of textual data mining is a topic field that increases in usefulness and interest over time [46].

Trend mining and discovery is useful in many application domains such as: market monitoring, medical diagnosis, opinion mining, stock market analysis, etc. and due to the increase in the availability of text data on the Internet, trend mining research is becoming more and more research area [45].

3.2.6. Opinion Leaders

Opinion leaders are the enlighten small number of people in a community that intercept information from mass media, interpret the information, and diffuse the information they receive to the personal networks that they themselves belong [3]. Opinion leaders can be politicians, business leaders, community leaders, journalists, educators, celebrities and sports stars [47].

There are opinion reviews for the products in amazon.com, and the reviews have been organized in order to give customers chance to view and read about the perception of other customers that used the products. The comment can help someone to decide on which product is better. But the system in amazon.com could not summarize the features that the previous customers described unless the customer read all of the comments which can consume a great amount of time. With the help of opinion mining and text summarization, it is possible to summarize all the features and customers perception on a product.

There are three important entities in opinion leaders, these entities are (a) the target entity on which opinion is expressed, (b) an author or the opinion leader who expressed the opinion, and (c) feeling or sentiment about the entity held by the opinion leader or the author [48]. Some of the opinion leader detection systems use sentiment analysis techniques and some use network analysis techniques; the details were given in the Related Work chapter of this study.

3.3. Analysis of Errors in Text Mining

In all the process of the classification, clustering and other text mining tasks, there are accuracies and errors. Understanding errors can give insight on how an algorithm or a program performs, so it is essential to understand the different error types. The errors and accuracies are as follows:

3.3.1. Overfitting

Overfitting occurs in a situation whereby an algorithm captures the noise of the data. Overfitting is often a result of an excessively complicated model and it can be a worse algorithm for another data [18].

3.3.2. Underfitting

Underfitting occurs when an algorithm cannot capture the underlying trend of the whole data. Underlying is often a result of an excessively simple model [18].

3.3.3. Accuracy

The accuracy is the ratio of the documents correctly categorized to the total number of documents [49].

$$Accuracy = \frac{Documents\ Correctly\ Categorized}{Total\ Documents} \quad \textbf{(3.1)}$$

3.3.4. Recall

Recall refers to the proportion of the total number of relevant results obtained to the total number of relevant results in the database [49].

A: $Number\ of\ relevant\ results\ retrieved.$

B: $Number\ of\ relevant\ results\ not\ obtained.$

$$\text{Recall} = \frac{A}{A+B} \quad \textbf{(3.2)}$$

3.3.5. Precision

Precision refers to the proportion of the number of relevant results obtained to the total relevant and irrelevant result obtained [49].

X: $Number\ of\ irrelevant\ results\ obtained.$

Y: $Number\ of\ relevant\ results\ obtained.$

$$\text{Precision } = \frac{Y}{X+Y} \quad \textbf{(3.3)}$$

The performance of the program is calculated and discussed in the Results and Discussion section of this thesis.

4. APPLICATION FOR FINDING OPINION LEADERS ON TWITTER

This chapter aims to give the general structure of the methods which are proposed in this study and the used processes during the findings and the analyses of the opinion leaders on Twitter. The thesis is based on general opinion leadership analysis which is not a domain specific. There are two types of data used in this thesis; the first one is the data retrieved with the help of Twitter Streaming API. The data is used for determining opinion leadership based on indegree, retweets, and sentiment influences. The details are given in the below paragraph. The second dataset used in this thesis is prepared by Go, Bhayani, and Huang (2009), the dataset contains both training and testing tweets data, but only their training dataset is used for training the sentiment classifier in this study. This dataset includes 798,622 negative tweets and 249,953 positive tweets [50]. The training data helps in classifying the retrieved 5,355 tweets through the Twitter Streaming API into positive or negative. Figure 4.1 below gives the representation of the processes and methods which are used in this study.

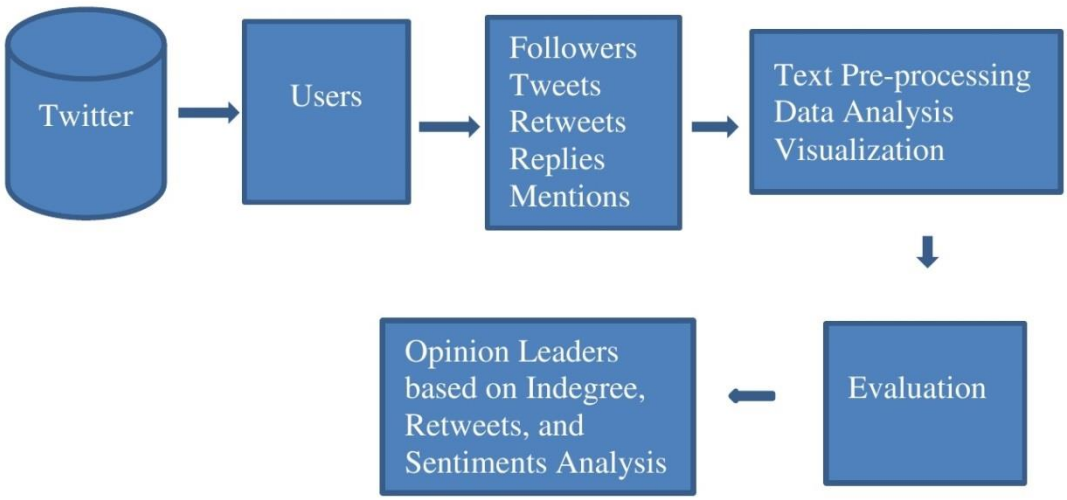

Figure 4.1. Opinion Leader Analysis Steps

The above Figure 4.1 represents the whole procedure of the analyses of the opinion leaders. The figure shows the platform, users, tweets data extraction, text pre-processing, data analysis, visualization and final analysis of the opinion leaders based on indegree, retweets, and sentiment analysis.

Spyder IDE has been used as a tool for coding and data collection. The code was entirely written in Python Programming Language. Python was chosen because there are a lot of tutorials and libraries for big data projects.

Table 4.1. Selected Candidate Opinion Leaders for test.

No	User
1	@nokia
2	@Cristiano
3	@Oprah
4	@nytimes
5	@CNN
6	@NASA
7	@SamsungMobile
8	@FIFAcom
9	@realDonaldTrump
10	@SupportSyrians
11	@HillaryClinton
12	@Number10gov
13	@BarackObama
14	@SenSanders
15	@SaharaReporters
16	@MBuhari
17	@TheDemocrats
18	@APCNigeria
19	@daily_trust
20	@HouseGOP
21	@mehdirhasan
22	@AlikoDangote
23	@ajplus
24	@PdpNigeria
25	@GarShehu
26	@alinuhu
27	@Arab_News
28	@aminugamawa
29	@GovAyoFayose
30	@Yadomah
31	@GEJonathan
32	@DrIsaPantami
33	@UN

Table 4.1 above presents the 33 candidate opinion leaders tested with this application on Twitter virtual community. The list above only gives the usernames of the tested users. The users are not listed based on any rank so far. All the three methods for opinion leader

discovery are tested using the same Twitter users, the description of the three methods and the approaches are shown in the following subheadings.

The application is implemented and tested with the data that is collected from randomly selected 33 Twitter users or candidate opinion leaders. The data collection from the candidate opinion leaders takes place from February 28, 2016 to March 16, 2016. Firstly, a benchmark was set from February 28, 2016 i.e. the day the data collections commence. The benchmark states that data from some Twitter users with at least 5000 followers will be retrieved for the analysis of the opinion leadership. The number of followers was set to 5000 because enough data may not be available from a user with less number of followers.

4.1. Opinion Leaders Based on Indegree

Twitter and many other networking sites have ranked Twitter users based on their number of followers (also called indegree). So, nowadays a user in the Twitter virtual community is usually considered influential if he or she has a large number of followers. Similar studies which consider the number of followers as a measure of influence has been reported in [25] and [29].

4.1.1. Approach

The approach towards the indegree ranking is just the traditional Twitter measure of influence i.e. ranking users based on the number of followers they have. The detailed data of the selected users to be tested are retrieved through the Twitter Streaming API. Most of the details are not needed for this study. The data needed in this thesis has been provided.

4.1.2. Algorithmic Approach

The algorithmic approach is not very complicated. After specifying which users to analyze, their data are retrieved with the Twitter Streaming API. The data are filtered and then a famous sorting algorithm known as Bubble Sort algorithm is used to sort the users and view the result in ascending order. The users are ranked according to the user with the highest indegree (or number of followers) to the user with the lowest number of followers. Bubble Sort algorithm is known to be a bit slow due to a number of swap and element comparison, but it was used because it is easy be implemented in such applications. Moreover, the sorting section is not dependent on any sorting algorithm.

4.1.3. Indegree Ranking Result

The below Table 4.2 gives the list of the 33 users in descending order i.e. from the user with the highest number of followers to the user with the lowest number of followers. It should be noted that the data were collected on March 16, 2016. So, the data may not reflect the current status of the users in the Twitter platform because the data is almost changing in each and every second.

Table 4.2. Indegree Ranking

RANK	USERNAME	TWEETS	FOLLOWING	FOLLOWERS
1	@BarackObama	14,726	637,233	71,121,116
2	@Cristiano	2,510	93	41,067,572
3	@Oprah	11,015	247	31,793,503
4	@nytimes	223,042	986	25,474,116
5	@CNN	81,165	1,115	23,910,145
6	@NASA	41,414	256	15,207,783
7	@SamsungMobile	3,935	432	11,589,192
8	@FIFAcom	52,471	790	8,292,576
9	@realDonaldTrump	31,322	42	6,952,963
10	@UN	44,547	1,064	6,791,451
11	@HillaryClinton	4,756	641	5,704,020
12	@Number10gov	10,464	735	4,317,464
13	@nokia	38,828	165,201	2,102,495
14	@SenSanders	13,892	1,928	1,691,642
15	@SaharaReporters	45,243	597	946,435
16	@MBuhari	1,672	51	524,433
17	@TheDemocrats	17,255	1,106	481,519
18	@APCNigeria	26,044	171	382,047
19	@daily_trust	143,829	143	321,715
20	@HouseGOP	9,575	10,215	280,173
21	@mehdirhasan	84,862	1,523	235,520
22	@AlikoDangote	64	17	195,772
23	@ajplus	24,528	318	195,483
24	@PdpNigeria	8,093	430	118,221
25	@GarShehu	1,865	93	112,931
26	@alinuhu	18,599	1,727	88,154
27	@Arab_News	67,528	255	71,413
28	@aminugamawa	22,988	2,409	39,217
29	@GovAyoFayose	1,966	91	35,100
30	@Yadomah	68,280	990	34,948
31	@GEJonathan	63	7	18,467
32	@DrIsaPantami	699	251	5,343
33	@SupportSyrians	269	510	5,028

Table 4.2 above describes the detailed data of the tested users. The data from column one is as follows: user rank, username, number of tweets, number of friends (people that the user follows), and the number of followers which is the number used in the ranking process.

4.2. Opinion Leaders Based On Retweet Influence

There are a number of researches that attempted to determine the influence of Twitter users based on the total number of retweets they received [29]. Retweets influence shows how important is the content of the tweets of a user. If many followers or users retweet some tweets or a single tweet of a user that mean those users are greatly influenced by the content of the tweets of that user. So, having a large number of retweets always show how influential a user is in the Twitter virtual community. Retweeting is simply saying that I agree with what this user says and you should also have a look at his or her tweet. The ranking in this chapter of the study is based on the tweets of a user (recent 100 tweets) that reach the highest number of people. In other words, a user is considered to be the most influential if recent 100 of his or her tweets have reached the highest number of people via retweets and then followed by the one that received less number in order and so on. The user in the number one position is the one with the highest number of audience reach through retweets. Thus, that user is the most influential in terms of retweets.

4.2.1. Approach

The approach for ranking users based on retweets is done with the help of an intermediate tool called tweetreach.com. The application estimates how many users did recent 100 tweets of a user reach through retweets. Thus, for all the selected twitter users (candidates opinion leaders), the numbers of account reach by their recent 100 tweets were determined. As an illustration, the way the ranking was done is by searching @username for a user. The same query was applied for all the candidate opinion leaders selected.

4.2.2. Algorithmic Approach

The algorithmic step comprises several steps. It starts with collecting the users to be analyzed (i.e. candidate opinion leaders), and then getting the retweets influence score, then sorting the users as follows:

- The first step is getting all the candidate opinion leaders to be analyzed.

- The second step is getting the number of accounts or people reach per 100 of the tweets of each of the candidate opinion leaders through retweets. As mention above, how many Twitter accounts recent 100 tweets of a user reach through retweets is collected with an intermediate tool called tweetreach.com.
- The third step is sorting the number of people each candidate opinion leader reach with his 100 tweets. Bubble sort algorithm is used for the sorting purpose.

4.2.3. Finding Retweets Influence

The retweets influence is found by sending a query with the @username of a Twitter user to be analyzed. For example, the two screenshots in Figures 4.2 and 4.3 below give the sample of query and the result obtain for @MBuhari.

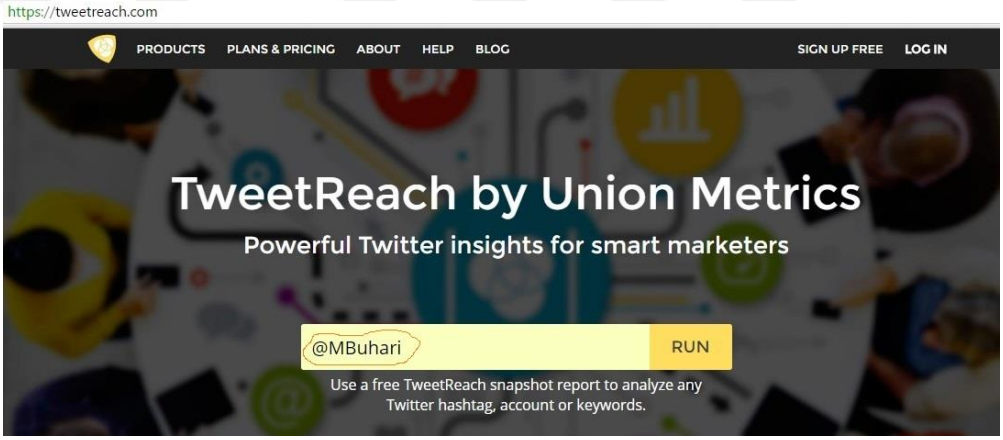

Figure 4.2. Retweet ranking query example

The figure above displays the screenshot of how to send the query. So far, there is no interface to integrate the application with Python or any other programming language. As a result, the queries are sent manually for all the candidate opinion leaders.

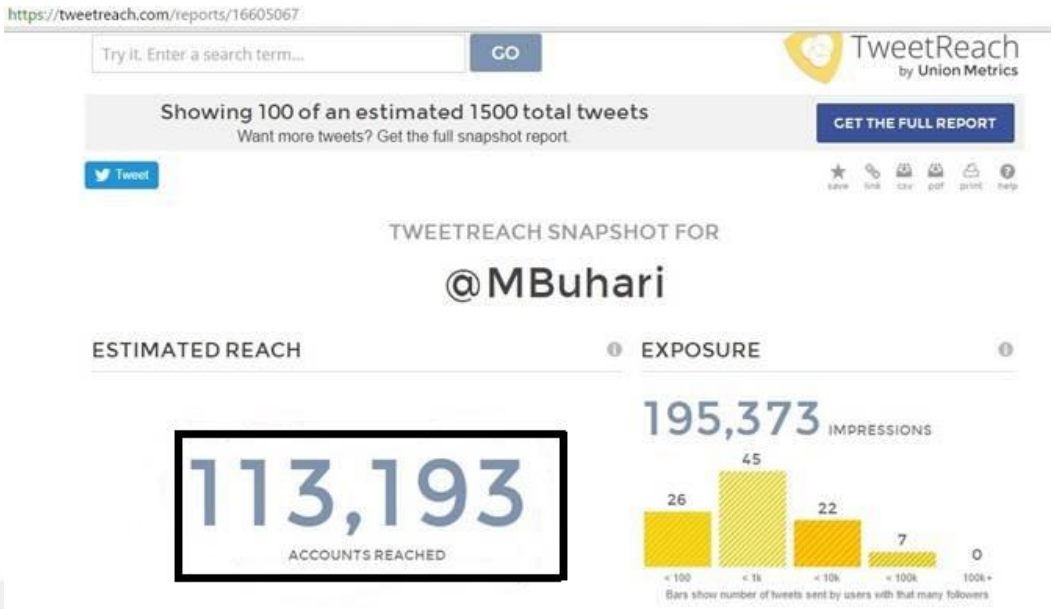

Figure 4.3. Retweet ranking query result

The above Figure 4.3 gives the sample output of the query with number of account reach per 100 recent tweets of this user. The recent 100 tweets of @MBuhari have reached 113,193 different accounts on the Twitter virtual community.

4.2.4. Retweets Ranking Result

The below Table 4.3 gives the list of the 33 users tested with this application in descending order i.e. from the user with the highest number of retweets score to the user with the lowest number of retweets score. It should be noted that the data were collected on March 14, 2016 as mentioned earlier. This means that the degree of retweets influence of a user may increase or decrease depending on the number of retweets received at the time of data collection.

Table 4.3. Retweets Ranking

RANK	USERNAME	RETWEET SCORE
1	@SamsungMobile	2,172,911
2	@Arab_News	1,596,023
3	@FIFAcom	1,083,230
4	@PdpNigeria	791,410
5	@BarackObama	610,598
6	@GovAyoFayose	414,050
7	@Cristiano	391,665
8	@Mehdirhasan	295,106
9	@Nokia	266,787
10	@SupportSyrians	264,145
11	@Oprah	262,004
12	@Yadomah	170,534
13	@Nytimes	165,267
14	@alinuhu	162,423
15	@TheDemocrats	156,377
16	@GEJonathan	144,190
17	@UN	142,831
18	@HouseGOP	132,753
19	@AlikoDangote	129,965
20	@HillaryClinton	125,839
21	@NASA	115,729
22	@MBuhari	109,170
23	@GarShehu	108,090
24	@realDonaldTrump	88,490
25	@Number10gov	87,336
26	@CNN	86,607
27	@daily_trust	81,866
28	@SaharaReporters	69,345
29	@ajplus	66,844
30	@APCNigeria	56,771
31	@SenSanders	42,307
32	@aminugamawa	41,532
33	@DrIsaPantami	14,670

The user with the highest retweets rank is @SamsungMobile. The retweets score of this user is 2,172,911. The retweet score of this user is very high, so, it is expected that some other influential Twitter users have retweeted the tweets of this user.

4.3. Opinion Leaders Based on Tweets Sentiments

There is no much work on ranking opinion leaders based sentiments of the total mentions they received in the virtual community. One of the studies which consider sentiment for opinion leadership was reported by Wu, Li, et al (2015). The authors stated that a candidate opinion leader is not a real opinion leader if most of the feedback or comments he or she received are negatives [51]. This chapter of the proposed study describes the steps and the techniques which are used in ranking candidate opinion leaders based on the sentiments of mentions (mentions here includes: replies and mentions on the Twitter) they received from other users on Twitter.

There are three important entities that are essential for better understanding of how ranking based on sentiment analysis works. The target entities are as follows:

- The entity on which the opinion was expressed; here the target entity is the candidate opinion leader whom the authors of tweets expressed their sentiment on him with their tweets.
- An author, who expressed the opinion. The author who expressed the opinion can be opinion leader or normal individual. This study does not consider who made the comment or who expressed the opinion rather is considers the opinion that was expressed.
- Feeling or sentiment about the entity (candidate opinion leader) held by the author of the comment (tweet mentions in this case). The sentiments of the mentions each candidate opinion leader received were used in the ranking based on binary classification as positive or negative.

Total of 5,355 Tweets feedback were retrieved from the candidate of opinion leaders. These tweets feedback are the mentions, replies, and every tweet containing a Twitter username of a user that is written as @username. The sentiments of these data were used for the ranking of the opinion leaders based on sentiments analysis.

The 5,355 tweets data retrieved from the candidate opinion leaders were labeled manually. The tweets were labeled manually for the purpose of determining the efficiency of the program. The labeling is as follows:

- Texts that contain positive sentiment were labeled as 1 (1 represents the positive class).

- Texts that contain negative sentiment were labeled as 0 (0 represents the negative class).

4.3.1. Approach

There are five steps in the approach of finding the opinion leaders based sentiments analysis, these steps are as follows:

- Candidate opinion leaders were selected.
- All the available mentions (replies and mentions) of every candidate user that are available on Twitter were collected.
- The mentions for every candidate user were collected separately.
- There were a total of 5,355 mentions for the tested candidate opinion leaders as mentioned earlier.
- A file is dedicated for mentions received by each user, thus, there were 33 different text files for all the candidate opinion leaders.

4.3.2. Algorithmic Approach

The algorithmic approach starts with the text collection and preprocessing. An alternative of Hash Map called counter was used to store the list of words after reading the tweets from text files. The counter is a dictionary subclass for counting hash table objects in Python. The counter class was selected because it is comfortable to implement in Python Programming Language.

While calculating the probability, logarithms of the result were added instead of multiplying the probability of the result. Logarithms were used so that 0 probabilities problem is eliminated. The 0 probability problem will make the result 0 if not eliminated as described with mathematical terms in Text Mining chapter. All the words were converted to lower case letters. Finally, Bubble sort algorithm was used to sort the sentiment scores in descending order.

4.3.3. Finding Tweets Sentiments with Naive Bayes

This section aims to give the general mathematical formula and the idea behind the text classification with the Naïve Bayes. An example of text classification was given in order to provide the reader with a better understanding of the classification process. The methods and procedures of classification of the text (Positive and Negative Texts) are divided into

two classes (positive class or negative class) were presented with detailed explanation of the approach.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \textbf{(4.1)}$$

The equation (4.1) above originated from Reverend Thomas Bayes [52]. It is a theorem of probability that is widely used in machine learning for classification purpose. The Bayesian Learning for text classification as reported by Mitchell (1997) is described by the following equations [53]. In this thesis, multivariable version of the equation was used to classify the tweets of users and it is written as follows:

$$P(positive|x_1, x_2, x_3, \dots x_n) = \frac{P(x_1, x_2, x_3, \dots x_n|positive)P(positive)}{P(x_1, x_2, x_3, \dots x_n)} \quad \textbf{(4.2)}$$

<>

$$P(negative|x_1, x_2, x_3, \dots x_n) = \frac{P(x_1, x_2, x_3, \dots x_n|negative)P(negative)}{P(x_1, x_2, x_3, \dots x_n)} \quad \textbf{(4.3)}$$

where $x_1, x_2, x_3, \dots x_n$ are the words that appear in the tweets.

Because $P(x_1, x_2, x_3, \dots x_n)$ is common in both classes it is eliminated and the equation becomes:

$$P(x_1, x_2, \dots x_n|positive)P(positive) <> P(x_1, x_2, \dots x_n|negative)P(negative)$$

The priori distribution for the two classes was assumed to be constant for the classification. So, there is no difference between $P(positive)$ and $P(negative)$. This means that ML and MAP are the same here.

After expansion, the equation will be equal to the below equation:

$$P(x_1 \,|positive)P(x_2 \,|positive) \dots P(x_n \,|positive) \quad \textbf{(4.4)}$$

<>

$$P(x_1 \,|negative)P(x_2 \,|negative) \dots P(x_n \,|negative) \quad \textbf{(4.5)}$$

If for example, substituting x_i=0 in one of the equation, then everything becomes zero like $P(C)\prod_{i=1}^{N} P(x_i \,|C) = 0$. So, to avoid 0 problem that will make everything zero, log was taken as described below.

$$log(P(x_1 \,/\, positive)) + log(P(x_2 \,/\, positive)) + \dots + log(P(\mathrm{x_n} \,/\, positive)) \quad \textbf{(4.6)}$$

<>

$$log(P(x_1\ /negative)) + log(P(x_2\ /negative)) + \ldots + log(P(\mathrm{x_n}\ /negative)) \quad \textbf{(4.7)}$$

Finally, the class with the highest value is selected as the class of the tweets i.e. between equations (4.6) for positive sentiment and (4.7) for negative sentiment. The Naive Bayes with Python programming were used to classify tweets written in English Language. The tweets samples were classified as either positive or negative.

4.3.3.1. Finding Tweets Sentiments Example

This part is dedicated for illustrating how the Naïve Bayes Classifier classifies the tweets sentiments from the scratch. The example is described by Zhang (2006) as follows: There are two classes, positive class (pos) and negative class (neg), then there are training data sample and test data sample [54].

Two Classes: Positive Class (pos) and Negative Class (neg)

Training Data

pos *d*1: "good."

pos *d*2: "very good."

neg *d*3: "bad."

neg *d*4: "very bad."

neg *d*5: "very bad, very bad."

Test Data

? *d*6: "good? bad! very bad!"

Naïve Bayes Learning

Prior Probability

P(pos) = *N*pos / (*N*pos+*N*neg) = 2 / (2+3) = 0.40

P(neg) = *N*neg / (*N*pos+*N*neg) = 3 / (2+3) = 0.60

// Constructing the Unigram Language Model

The Add-One Smoothing technique

P(very|pos)

= (*T*very|pos+1) / ((*T*very|pos+1) + (*T*good|pos+1) + (*T*bad|pos+1))

= (1+1) / ((1+1) + (2+1) + (0+1))

= 2 / (2 + 3 + 1)

= 0.33

P(good|pos)

= (Tgood|pos+1) / ((Tvery|pos+1) + (Tgood|pos+1) + (Tbad|pos+1))

= (2+1) / ((1+1) + (2+1) + (0+1))

= 3 / (2 + 3 + 1)

= 0.50

P(bad|pos)

= (Tbad|pos+1) / ((Tvery|pos+1) + (Tgood|pos+1) + (Tbad|pos+1))

= (0+1) / ((1+1) + (2+1) + (0+1))

= 1 / (2 + 3 + 1)

= 0.17

P(very|neg)

= (Tvery|neg+1) / ((Tvery|neg+1) + (Tgood|neg+1) + (Tbad|neg+1))

= (3+1) / ((3+1) + (0+1) + (4+1))

= 4 / (4 + 1 + 5)

= 0.40

P(good|neg)

= (Tgood|neg+1) / ((Tvery|neg+1) + (Tgood|neg+1) + (Tbad|neg+1))

= (0+1) / ((3+1) + (0+1) + (4+1))

= 1 / (4 + 1 + 5)

= 0.10

P(bad|neg)

= (Tbad|neg+1) / ((Tvery|neg+1) + (Tgood|neg+1) + (Tbad|neg+1))

= (4+1) / ((3+1) + (0+1) + (4+1))

= 5 / (4 + 1 + 5)

= 0.50

Naïve Bayes Classification

The Likelihood

The Application of the Unigram Language Model for the classification.

P(d6|pos)

= P(good|pos)×P(bad|pos)×P(very|pos)×P(bad|pos)

= 0.50×0.17×0.33×0.17

= 0.0048

P(d6|neg)

= P(good|neg)×P(bad|neg)×P(very|neg)×P(bad|neg)

= 0.10×0.50×0.40×0.50

= 0.010

Posterior Probability

P(pos| *d*6)

= P(*d*6|pos)×P(pos)/P(*d*6) // Bayes' Rule

= 0.0048×0.40/P(*d*6)

= 0.0019/P(*d*6)

P(neg| *d*6)

= P(*d*6|neg)×P(neg)/P(*d*6) // Bayes' Rule

= 0.010×0.60/P(*d*6)

= 0.0060/P(*d*6)

The Classification

Since *P*(pos| *d*6) < *P*(neg| *d*6),

*d*6 should be classified into the neg class.

P.S.

Note that without the Add-One (Laplace) smoothing,

P(good|neg) would be 0 and consequently P(neg| *d*6) would be 0.

The Add-One (Laplace) smoothing does the same task as taking the logarithm of the multiplication, because if P(neg| d6) > P(pos=g| d6) also log(P(neg| d6)) > (P(pos=g| d6)). So, the logarithms technique was used to solve the 0 problem in this study.

4.3.4. Sentiments Ranking Result

The sentiment score is calculated using the following variables and formulas:

$P: is\ the\ total\ number\ of\ positive\ mentions\ received\ by\ a\ user.$

$N: is\ the\ total\ number\ of\ negative\ mentions\ received\ by\ a\ user.$

$$positive = \left(\frac{P}{P+N}\right) X100 \quad \textbf{(4.8)}$$

$$negative = \left(\frac{N}{P+N}\right) X100 \quad \textbf{(4.9)}$$

For example, the user called @NASA has 82 value of *P* also called positive tweets mentions and 93 value of *N* also called negative tweets mentions. When the formula is

applied @NASA has a negative sentiment score of 53.1428571429% and a positive sentiment score of 46.8571428571%.

Table 4.4 below gives the list of the tested 33 candidate opinion leaders and their sentiments scores in descending order i.e. from the user with the highest positive sentiments received to the user with the lowest positive sentiments received. Both the negative and positive sentiments scores for all the users are given, the negative scores in the third column and the positive scores in the fourth column of the table.

Table 4.4. Tweets Mentions Sentiment Ranking

Rank	Username	Negative (%)	Positive (%)
1	@DrIsaPantami	34.8837209302	65.1162790698
2	@Oprah	36.6515837104	63.3484162896
3	@UN	50.3875968992	49.6124031008
4	@Number10gov	50.5576208178	49.4423791822
5	@NASA	53.1428571429	46.8571428571
6	@alinuhu	53.2608695652	46.7391304348
7	@AlikoDangote	53.5545023697	46.4454976303
8	@SupportSyrians	53.9682539683	46.0317460317
9	@BarackObama	54.0540540541	45.9459459459
10	@mehdirhasan	56.3291139241	43.6708860759
11	@GEJonathan	57.0000000000	43.0000000000
12	@TheDemocrats	57.6612903226	42.3387096774
13	@nokia	58.7064676617	41.2935323383
14	@HillaryClinton	59.7902097902	40.2097902098
15	@HouseGOP	60.3448275862	39.6551724138
16	@realDonaldTrump	62.7177700348	37.2822299652
17	@GovAyoFayose	63.2850241546	36.7149758454
18	@SenSanders	63.3136094675	36.6863905325
19	@PdpNigeria	63.4730538922	36.5269461078
20	@daily_trust	65.3846153846	34.6153846154
21	@Yadomah	66.1538461538	33.8461538462
22	@ajplus	67.6923076923	32.3076923077
23	@MBuhari	69.5067264574	30.4932735426
24	@aminugamawa	69.6969696970	30.3030303030
25	@Arab_News	69.8630136986	30.1369863014
26	@GarShehu	70.6161137441	29.3838862559
27	@APCNigeria	73.3067729084	26.6932270916
28	@CNN	73.6842105263	26.3157894737
29	@SamsungMobile	74.0259740260	25.9740259740
30	@Cristiano	74.3750000000	25.6250000000
31	@nytimes	75.9259259259	24.0740740741
32	@FIFAcom	76.1904761905	23.8095238095
33	@SaharaReporters	79.7619047619	20.2380952381

The analysis and all the discussions of all the methods and approaches used in the ranking process are given in the results and discussion chapter of this study.

5. RESULTS AND DISCUSSION

The main aim of this research is to analyze opinion leaders (or influential leaders) using text mining techniques on social media. The research investigates the influence of some specific Twitter users on the Twitter microblogging site. Prior researches on the opinion leadership analysis that were reported in the literature are mainly concentrated on domain specific opinion leadership analyses. However, this research focuses on general opinion leadership which is not dependent on any specific domain.

This research uses three different approaches for the analysis of the influential users on the Twitter platform. These approaches are:

- Indegree: Indegree here means the total number of followers a user has. Traditionally, a user with a high number of followers is generally considered as an influential individual.
- Retweets: Retweet refers to the reposting of the tweet(s) of another user. Retweet usually shows agreement by sharing the exact Tweet(s) someone else posted.
- Sentiments of Mentions: Mention is any Tweet that contains the Twitter username of a user starting with @; just like this: "Hello @username, that's cool. Very kind of you". Mentions are divided into two types: Mentions with the username placed in the beginning of the tweet and mentions the username placed in the middle of the text. The sentiments expressed in the texts or sentences via the tweets mentions were used for the opinion leadership analysis.

5.1. Indegree Influence Analysis

Analysis of the results based on indegree or number of followers among the selected 33 Twitter users suggests the following:

- The most influential user is @BarackObama. What does this mean? It means that when the number of followers is considered as the criteria for the determination of the opinion leadership, it can be categorically said that @BarackObama is the most influential user because the number of followers he has at the time of the data collection is 71,121,116. This number of his followers is greater than the number of followers of each user among the remaining 32 analyzed users

- The less influential user is @SupportSyrians. This user is the less influential because the user has 5,028 followers and this number is the smallest among the other 32 users.
- The top ten most influential users based on indegree ranking are as follows:
    1. @BarackObama
    2. @Cristiano
    3. @Oprah
    4. @nytimes
    5. @CNN
    6. @NASA
    7. @SamsungMobile
    8. @FIFAcom
    9. @realDonaldTrump
    10. @UN

The ten influentials above starts from the most influential to the one that follows as numbered. In order words, the list above gives the top ten most influentials in descending order. The complete list of the 33 users with their influence scores in descending order was given chapter 4 of this thesis.

- The least ten most influential users based on indegree ranking are as follows:
    1. @PdpNigeria
    2. @GarShehu
    3. @alinuhu
    4. @Arab_News
    5. @aminugamawa
    6. @GovAyoFayose
    7. @Yadomah
    8. @GEJonathan
    9. @DrIsaPantami
    10. @SupportSyrians

The list above gives the least ten influentials among the 33 selected users. The user in the number 10 position here is the least influential among all the 33 users.

5.2. Retweets Influence Analysis

Analysis of the results based on retweets influences among the selected 33 Twitter users suggest the following findings:

- The most influential user is @SamsungMobile. It means that when retweets influence is considered as the criteria for the determination of the opinion leadership, it can be categorically said that @SamsungMobile is the most influential user. The retweets influence score of this user at the time of the data collection is 2,172,911. Thus, the retweets influence score of this user is greater than the score of each other user among the remaining 32 users.
- The least influential user is @DrIsaPantami. This user is the least influential because the user has 14,670 retweets influence score and this number is the smallest among the other 32 users.
- One of the important findings here revealed that having large followers (i.e. high indegree rank) does not necessarily mean a user is influential. Because the findings based on retweets show that some users with less number of followers score higher rank in comparison with those that score higher with indegree ranking. This finding is particularly important for us because there are some works in the literature that support this finding. For example, Cha, Haddaddi, Benevenuto and Gummadi (2010), reported that famous users with high indegree may not necessarily be influential in terms of mentions or retweets [30]; this finding is exactly as the finding in this thesis.
- The best ten most influential users based on retweets ranking are as follows:
    1. @SamsungMobile
    2. @Arab_News
    3. @FIFAcom
    4. @PdpNigeria
    5. @BarackObama
    6. @GovAyoFayose
    7. @Cristiano
    8. @Mehdirhasan
    9. @Nokia
    10. @SupportSyrians

The 10 influentials above start from the most influential based on retweets to the one that follows as numbered. In order words, the list above gives the top ten most influentials based on retweet ranking in descending order. The complete list of the 33 users with their influence scores in descending order was given in the chapter 4.

There is an important observation about the user that is the most influential according to retweets influence score, that is @SamsungMobile. Even though there is no any proof for this claim, but it is expected that this user (@SamsungMobile) has reached to other influentials leaders on the Twitter platform so that they post and advertise to the media community about Samsung mobile device. Thus, this makes the user very popular in terms of spreading the news about the user via retweets.

- The least ten most influential users based on retweets ranking are as follows:
    1. @realDonaldTrump
    2. @Number10gov
    3. @CNN
    4. @daily_trust
    5. @SaharaReporters
    6. @ajplus
    7. @APCNigeria
    8. @SenSanders
    9. @aminugamawa
    10. @DrIsaPantami

The user in the number 10 position here is the least influential among all the 33 users in terms of retweets ranking. In order words, the user named @DrIsaPantami is the number 33 in terms of retweets influence.

- Retweets influence is a measure of agreement by what the author of tweet posted. Having large number retweets show that a user is influential in the Twitter virtual community.

5.3. Sentiment Influence Analysis

Analysis of the results based on sentiments influence score among the selected 33 Twitter users revealed the following findings:

- The most influential user is @DrIsaPantami. The sentiment influence score of this user at the time of the data collection is 65.1162790698% which is the highest score among all the other users.
- The least influential user is @SaharaReporters. The sentiment influence score of this user at the time of the data collection is 20.2380952381% which is the smallest score among all the other users.
- The best ten most influential users based on sentiments ranking are as follows:
    1. @DrIsaPantami
    2. @Oprah
    3. @UN
    4. @Number10gov
    5. @NASA
    6. @alinuhu
    7. @AlikoDangote
    8. @SupportSyrians
    9. @BarackObama
    10. @Mehdirhasan

The list above gives the top ten most influentials based on sentiments ranking in descending order. The complete list of the 33 users with their influence scores in descending order was given in the chapter 4. The result was found using equations (4.6), (4.7), (4.8) and (4.9) as explained in chapter 4 section 4.3.

- The least ten most influential users based on sentiment ranking are as follows:
    1. @aminugamawa
    2. @Arab_News
    3. @GarShehu
    4. @APCNigeria
    5. @CNN
    6. @SamsungMobile
    7. @Cristiano
    8. @nytimes
    9. @FIFAcom
    10. @SaharaReporters

- The ten users above are the least influential among all the 33 users in terms of sentiments ranking.
- The sentiments influence criteria also revealed that having a large number of followers does not necessary mean that a user is influential. A user may have a large number of followers, but that does not mean the majority of the followers agree with him.

5.3.1. Accuracy Analysis

The performance of any developed program or algorithm is very important. So, to determine the accuracy of the sentiments classification task used in the tweets sentiments ranking classifier, a confusion matrix was calculated. The manner in which the calculations take place is described by the following example: If there are a total of 100 tweets to be classified as positive or negative. If the first 50 tweets are positive and the second 50 tweets negative, but the program classified the first 50 positive tweets as: 40 positive and 10 negative, and the second 50 negative tweets as: 30 negative and 20 positive. This means that based on this example there are 40 True Positive (TP), 10 False Negative (FN), 30 True Negative (TN), and 20 False Negative (FN). The following table, Table 5.1, gives the result of the classification of 5,355 tweets in this study as follows:

Table 5.1. Confusion Matrix

		Predicted Classes	
		Positive	Negative
Actual Classes	Positive	TP = 1078	FN = 1137
	Negative	FP = 1031	TN = 2109

The total accuracy, precision, and recall are calculated using equations (3.1), (3.2), and (3.3) in chapter 3 as shown below.

- $\text{Total Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{3187}{5355} = 59.51\%$
- $\text{Precision}_{\text{Positive}} = \frac{TP}{TP+FP} = \frac{1078}{1078+1031} = \frac{1078}{2109} = 51.11\%$
- $\text{Recall}_{\text{Positive}} = \frac{TP}{TP+FN} = \frac{1078}{1078+1137} = \frac{1078}{2215} = 48.67\%$
- $\text{Precision}_{\text{Negative}} = \frac{TN}{TN+FN} = \frac{2109}{2109+1137} = \frac{2109}{3246} = 64.97\%$

- $\text{Recall}_{\text{Negative}} = \frac{TN}{TN+FP} = \frac{2109}{2109+1031} = \frac{2109}{3140} = 67.16\%$

The result of the Naïve Bayes sentiments classifier shows that the application has an accuracy of 59.51%. Another thing that can be learned from the confusion matrix table is that the classifier has higher accuracy when classifying negative tweets.

5.3.1.1. Classified Tweets Sample

This section of the study presents the samples of the tweets that were correctly classified and the samples of the tweets that were wrongly classified in the application. The tweets are in Table 5.2 below.

Table 5.2. Classified Tweets Sample

Class	Result	Tweet
1	TP	@DrIsaPantami May the Peace, Mercy &Blessing of Allah be upon you all.
1	FN	@DrIsaPantami Is my pleasure sir. Best wishes
1	TP	@DrIsaPantami Indeed it is always good to be good. Only good comes out from the good and bad usually emanates
0	TN	@AlikoDangote govt policies is designed 2 hurting young epreneurs nd protecting well established
1	TP	@AlikoDangote Good evening Mr Dangote, truly the future of our nations lies on our youth., and your intiatives are welcome in zambia.Grace N
1	TP	@Cristiano you are amazing friend
1	TP	@SaharaReporters You need to activate the cell operations of ur reputable journalist company, starting with Kwara state.
1	FN	Was thinking..if only we had a kwara version of @SaharaReporters ...our people (kwarans) would be free of siege mentality...
0	FP	@inrnt that shit was nasty af. @Oprah you nasty af for making that movie. I had nightmares
1	TP	@Oprah @NinaShaw you should come to our BREATHE CONFERENCE ALL ABOUT EMPOWERMENT FOR WOMEN

The table above shows the original labels or classes of the tweets and classes that were determined by the developed application. The application counts all the tweets that were wrongly and correctly calculated separately. The result is used for the confusion matrix calculation.

5.3.2. Comparison of Accuracy

A comparison is necessary for researchers to determine the performance of their research and the reported researches in the literature. This subheading is dedicated for the sake of the comparison of the Naïve Bayes classifier used for the sentiments classification. In this study, the Naïve Bayes sentiment classifier obtained 59.51% accuracy as mentioned above.

A similar study for text polarity classification on Twitter was reported by Lima et al. (2015), their accuracy results [55] and the accuracy of the developed application are compared in Table 5.3 as shown below.

Table 5.3. Accuracy Comparison

ACCURACY %		
Author(s)	**Classifier**	**Result**
Developed Application	NB	59.51%
Lima et al. [55]	NB	63.011%
Lima et al. [55]	KNN	61.535%

Even though this thesis does not consider emoticons for the sentiment classification, there is no much difference in the accuracy of the reported study and the accuracy of the developed application.

5.4. Indegree, Retweets, and Sentiment Influence Analysis

There are quite interesting findings in the analysis of the opinion leaders on the Twitter social media site. Table 5.4 below lists some important details as follows:

Table 5.4. Opinion Leaders

INDEGREE & RETWEETS	INDEGREE & SENTIMENTS	RETWEETS & SENTIMENTS	ALL TIME INFLUENTIALS
@SamsungMobile	@Oprah	@SupportSyrians	@BarackObama
@Cristiano	@NASA	@BarackObama	
@FIFAcom	@UN	@Mehdirhasan	
@BarackObama	@BarackObama		

Table 5.4 above presents the list of the users that are influential based on the combination of two methods. There are 4 users that are influential based on Indegree and Retweets, 4 users are influential based on Indegree and Sentiments, and 3 users according to Retweets and Sentiments. The table also gives the all-time influential user, the user that is influential according to both the three methods.

- There are 11 users that are influentials according to two methods as shown in the table above.
- The user that happens to be influential according to all three methods is @BarackObama. Thus, this user is the most influential, the all-time influential leader.

An interesting example of how a user influences public was examined via the discovered all-time influential leader. The example is as follows:

There was one statement by @BarackObama during a speech, he said, "Muslim Americans are our friends and our neighbors, our co-workers, our sports heroes and yes, they are our men and women in uniform, who are willing to die in defense of our country." After this statement, for the first time in more than a year, the top Googled noun after "Muslim" was not "terrorists," "extremists" or "refugees." It was "athletes," followed by "soldiers." And, in fact, "athletes" kept the top spot for a full day afterward [56]. This is a good example of how influential individual affects the way people view things that are happening around them.

6. CONCLUSION

The study was set out to explore existing work on opinion leaders and proposed a reliable solution that will improve detection and analysis of opinion leaders on social media using text mining techniques. There are two approaches which researchers used to determine opinion leadership in the literature. The first approach is using network analysis; most of the researchers used this approach to determine the opinion leadership. The second approach is using text mining techniques; only a few researchers used text mining approach in the opinion leadership analysis. Text mining approaches were used to determine and analyze opinion leadership in this study.

The opinion leaders or influential leaders are the few individuals that many others refer to them for decision making. These opinion leaders can easily convince a large number of people in buying decisions, political campaigns, and many other issues via the words of mouth. The importance of opinion leaders in marketing made many companies change their advertising strategies towards identifying those opinion leaders so that the opinion leaders can influence or convince non-opinion leaders to adopt their products or services.

This research specifically determines and analyzes opinion leaders on Twitter social media platform using text mining techniques. The proposed study uses indegree, retweets, and sentiments of tweets mentions in the analysis process. The indegree technique used in this thesis determines the influential users considering their number of followers. Whereas the retweets technique determines the influential users considering how many other users did recent 100 tweets of a user reached via retweet; and the sentiments of mentions determines the influential by classifying those influential according to how many positive or negative comments they received from other Twitter users. Ranking based on sentiment score is one of the contributions of this research. The sentiment classification adopted for the ranking process uses Naïve Bayes binary classification method and it has an accuracy of 59.51%. The results have shown that retweets and sentiments of mentions are the real measures that determine the influence of users on the Twitter social media platform. The results have also shown that having large followers (i.e. high indegree rank) does not necessarily mean a user is influential based on retweets or sentiment of total mentions received.

There are certain limitations that might affect the performance of this study, these include:

- A number of account reach by recent 100 tweets of users was considered for retweets ranking. However, considering all tweets of users may give better performance for retweets ranking.
- Emoticons (such as: ☺,☹,:P,:], lol, (Y), etc.) were not considered in the sentiment classification process, yet such emoticons may express something about the sentiment of a comment. Thus, considering emoticons may improve the performance of the classification.
- Mentions received by users for some date interval were considered in the sentiment classification, but considering all mentions received by a user on the Twitter platform may give more reliable results.
- The relationship between indegree of users and the number of retweets received by users may be explored in the future.

The main contributions of this research are the analysis of the related studies on opinion leaders found in the literature, the determination of the opinion leaders on the Twitter social media platform using indegree, retweets, and sentiments analysis and the suggestions of the possible future study.

REFERENCES

[1] **Edward, K. and Berry, J.**, The Inflientials: One American In Ten Tells The Other Nine How To Vote, Where To Eat, And What To Buy., 2003.

[2] **Arndt, J.**, "Ole Of Product-Related Conversations In The Diffusion Of A New Product," Journal Of Marketing Research, Vol. 4, Pp. 291-295, 1967.

[3] **Katz, E. and Lazarsfeld, P.**, PERSONAL INFLUENCE: THE PART PLAYED BY PEOPLE IN THE FLOW OF MASS COMMUNICATIONS, 1955.

[4] **Chaudhry, S. A.**, "Opinion Leadership and Its Role In Buyer Decision Making," Academy Of Contemporary Research Journal, Vol. 2, No. 1, Pp. 16-23, 2013.

[5] **Agarwal, N., Liu H., Tang L. and Yu P. S.**, "Identifying The Influential Bloggers In A Community," Proceedings Of The International Conference On Web Search And Web Data Mining, 2008.

[6] **Drezner, D. W. and Farrell, H.**, "THE POWER AND POLITICS OF BLOGS," Http://Www.Utsc.Utoronto.Ca/~Farrell/Blogpaperfinal.Pdf, 2004.

[7] **Feldman, R. and Sanger, J.**, THE TEXT MINING HANDBOOK Advanced Approaches In Analyzing Unstructured Data, New York: CAMBRIDGE UNIVERSITY PRESS, 2007.

[8] **Binali, H., Potdar, V. and Wu, C.**, "A State Of The Art Opinion Mining And Its Application Domains," IEEE International Conference, 2009.

[9] **CHARNIAK, E. A. and Mcdermott, D. V.**, Introduction To Artificial Intelligence, Addison-Wesley Pub.Company, 1985.

[10] **Schalkoff, R. J.**, Artificial Intelligence: An Engineering Approach, Mcgraw-Hill, 1990.

[11] **Winston, P. H.**, Artificial Intelligence, Pearson, 1992.

[12] **Luger, G. F. and Stubblefield, W. A.**, Artificial Intelligence: Structures And Strategies For Complex Problem Solving, Benjamin/Cummings Pub. Co., 1993.

[13] **Mccarthy, J.**, "WHAT IS ARTIFICIAL INTELLIGENCE?," Http://Www-Formal.Stanford.Edu/Jmc/Whatisai/, 2007.

[14] **Murphy, K. P.**, Machine Learning A Probabilistic Perspective, London: The MIT Press.

[15] **Samuel, A. L.**, Some Studies In Machine Learning Using The Game Of Checkers, Https://Www.Cs.Virginia.Edu/~Evans/Greatworks/Samuel1959.Pdf, 1959.

[16] **Alpaydin, E.**, Introduction To Machine Learning, The MIT Press, 2004.

[17] **Gupta, H.**, Management Information System (An Inside), International Book House PVT. LTD., 2011.

[18] **Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P.**, "From Data Mining To Knowledge Discovery In Databases," American Association For Artificial Intelligence (AAAI), 1996.

[19] Twitter, "Types of Tweets and Where They Appear," Https://Support.Twitter.Com/Articles/119138, 2015.

[20] **Freimut, B. and Carolin, K.**, "Detecting Opinion Leaders and Trends In Online Communities," Fourth International Conference On Digital Society, 2010.

[21] **Wu, Y. and Ren, F.**, "Learning Sentimental Influence In Twitter," International Conference On Future Computer Sciences And Application, 2011.

[22] **Duan, J., Zeng, J. and Luo, B.**, "Identification Of Opinion Leaders Based On User Clustering And Sentiment Analysis," IEEE/WIC/ACM International Joint Conferences On Web Intelligence (WI) And Intelligent Agent Technologies (IAT), 2014.

[23] **Joshi, A., Finin, T., Java, A., Kale, A. and Kolari, P.**, "Web (2.0) Mining: Analyzing Social Media," Http://Ebiquity.Umbc.Edu/_File_Directory_/Papers/379.Pdf.

[24] **Zhou, H., Zeng, D. and Zhang, C.**, "Finding Leaders From Opinion Networks," IEEE, 2009.

[25] **Hui P. and Gregory M.**, "Quantifying Sentiment and Influence In Blogspaces," 1st Workshop On Social Media Analytics, 2010.

[26] **Merwe, R. V. D. and Heerden, G. V.**, "Finding And Utilizing Opinion Leaders: Social Networks And The Power Of Relationships," S.Afr.J.Bus.Manage.2009,40(3), 2009.

[27] **Papadakis, G. A., Tserpes, K. and Varvarigou, T.**, "Influence Patterns In Topic Communities Of Social Media," International Conference On Web Intelligence, Mining And Semantics, 2012.

[28] **Weng, J., Lim, E. P., Jiang, J. And He, Q.**, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," International Conference On Web Search and Data Mining, 2010.

[29] **Leavitt, A., Burchard, E., Fisher, D. and Gilbert, S.**, "The Influentials: New Approaches For Analyzing Influence On Twitter," A Publication Of The Web Ecology Project, 2009.

[30] **Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. P.**, "Measuring User Influence In Twitter: The Million Follower Fallacy," Association For The Advancement Of Artificial Intelligence, 2010.

[31] **Mustafaraj, E. and Metaxas, P. T.**, "What Edited Retweets Reveal About Online Political Discourse," Analyzing Microtext: Papers From The 2011 AAAI Workshop (WS-11-05), 2011.

[32] **Hung, C. and Yeh, P.-W.**, "Identification of Opinion Leaders Using Text Mining Technique In Virtual Community," Information Management And Big Data, 2014.

[33] **Anchez, D. S´, Mart´In-Bautista, M. and Blanco, I.**, "Text Knowledge Mining: An Alternative To Text Data Mining," IEEE International Conference On Data Mining Workshops, 2008.

[34] **Akaichi, J., Dhouioui, Z. and Pérez, M. J. L.-H.**, "Text Mining Facebook Status Updates For Sentiment Classification," IEEE, 2013.

[35] **Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F.**, "Predicting The Political Alignment Of Twitter Users," IEEE, 2011.

[36] **Mathiak, B. and Eckstein, S.**, "Five Steps To Text Mining In Biomedical Literature," Proceedings Of The Second European Workshop On Data Mining And Text Mining In Bioinformatics, 2004.

[37] **Gupta, V. and Lehal, G. S.**, "A Survey Of Text Mining Techniques And Applications," JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, Vols. VOL. 1, NO. 1, 2009.

[38**] Shi, G. and Kong, Y.**, "Advances In Theories And Applications Of Text Mining," The 1st International Conference On Information Science And Engineering (ICISE2009), 2009.

[39] **Pak, A. and Paroubek, P.**, "Twitter As A Corpus For Sentiment Analysis And Opinion Mining," In Proceedings Of The Seventh Conference On International Language Resources And Evaluation, 2010.

[40] **Farhadloo, M. and Rolland, E.**, "Multi-Class Sentiment Analysis With Clustering And Score Representation," 2013 IEEE 13th International Conference On Data Mining Workshops, 2013.

[41] Kunwar S., "Text Documents Clustering Using K-Means Clustering Algorithm.," Http://Www.Codeproject.Com/Articles/439890/Text-Documents-Clustering-Using-K-Means-Algorithm, 2013.

[42] **Martin, J.**, "Clustering Full Text Documents," IJCAI-95 Workshop On Data Engineering For Inductive Learning, 1995.

[43] **Sykes, A. O.**, "Introduction To Regression Analysis," Http://Www.Law.Uchicago.Edu/Files/Files/20.Sykes_.Regression.Pdf.

[44] **Radev, D. R., Hovy, E. and Mckeown, K.**, "Introduction To The Special Issue On Summarization," Mitpressjournals, Vol. 28, 2002.

[45] **Streibel, O. and Tolksdorf, R.**, "Mining Trends In Texts On The Web," Networked Information Systems, Free University, Berlin.

[46] **Kontostathis, A., Galitsky, L., Pottenger, W. M., Roy, S. and Phelps, D. J.**, "A Survey Of Emerging Trend Detection In Textual Data Mining," Http://Citeseerx.Ist.Psu.Edu/Viewdoc/Summary?Doi=10.1.1.19.4190, 2003.

[47] Grimsley, S., "Opinion-Leader In Marketing," Http://Study.Com/Academy/Lesson/Opinion-Leader-In-Marketing-Definition-Lesson-Quiz.Html.

[48] **Arora, R. and Srinivasa, S.**, "A Faceted Characterization Of The Opinion Mining Landscape," IEEE, 2014.

[49] **Junker, M., Hoch, R. and Dengel, A.**, "On The Evaluation Of Document Analysis Components By Recall, Precision, And Accuracy," IEEE Proceedings Of The Fifth

International Conference On Document Analysis And Recognition. ICDAR '99 (Cat. No.PR00318), Pp. 713 - 716, 1999.

[50] **Go, A., Bhayani, R. and Huang, L.**, "Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford 1, 12, 2009.

[51] **Wu, C., Li, C., Yan, W., Luo, Y., Mao, X., Du, S. and Li, M.**, "Identifying Opinion Leader In The Internet Forum," International Journal Of Hybrid Information Technology, Vol. 8, No. 11, Pp. 424-425, 2015.

[52] **Bellhouse, D. R.**, "The Reverend Thomas Bayes, FRS: A Biography To Celebrate The Tercentenary Of His Birth," Institute Of Mathematical Statistics, Vol. 19, No. 1, P. 3–43, 2004.

[53] **Mitchell, T.**, Machine Learning, Mcgraw Hill International Edition, 1997.

[54] **Zhang, D.**, "An Example Of Text Classification With Naïve Bayes," Http://Www.Dcs.Bbk.Ac.Uk/~Dell/Teaching/Ir/Examples/Nb_Example.Pdf, Birkbeck, University Of London, 2006.

[55] **Lima, A. C., Castro, L. D. and Rodríguez, J. M. C.**, "A Polarity Analysis Framework For Twitter Messages," Applied Mathematics And Computation 270, P. 756–767, 2015.

[56] **Soltas, E. and Stephens-Davidowitz, S.**, "The Rise Of Hate Search," Http://Www.Nytimes.Com/2015/12/13/Opinion/Sunday/The-Rise-Of-Hate-Search.Html?_R=1, 2015.

APPENDIX

Sample codes used in this thesis are given below.

```
/** * @author Kaloma_Usman_Majikumna
 */
def bayes(tweet, posMap, negMap):
    test = taskTest(tweet)
    posProb = 0.0
    negProb = 0.0
    global posTweets,negTweets
    # variables for percentage countin
    global TP,FP,TN,FN
#    posTweets = 0
#    negTweets = 0
    p_values=float(sum(posMap.values()))
    n_values=float(sum(negMap.values()))
    for word in test:
        #print                                                    'word:
',word,'posMap[word]',posMap[word],'float(sum(posMap.values()))',sum(posM
ap.values())
        posProb += log(posMap[word]/p_values)
        negProb += log(negMap[word]/n_values)
    if posProb > negProb:
        posTweets +=1
        #print "positive",
    else:
        negTweets +=1
        #print "negative",
    ################### Accuracy ####
    if ((tweet[0] == "1") and (posProb > negProb)):
        TP += 1
    if ((tweet[0] == "1") and (posProb < negProb)):
        FN += 1
    if ((tweet[0] == "0") and (negProb > posProb)):
        TN += 1
    if ((tweet[0] == "0") and (negProb < posProb)):
        FP += 1
#bayes ends

def getFollowers(dizi):
```

```
    #import tweepy
    auth = OAuthHandler(APP_KEY, APP_SECRET)
    auth.set_access_token(OAUTH_TOKEN, OAUTH_TOKEN_SECRET)
    api = tweepy.API(auth)

    obj = []
    for user in dizi:
        this = api.get_user(user)
#        print this.screen_name
#        print ' '
        username = this.screen_name
        followers = this.followers_count
        status = this.statuses_count
        followee = this.friends_count
        #print x,': Follwers::',y,': Tweets::',z
        data = [username,followers, status, followee]
        obj.append(data)
    return obj
#ends get followers

def taskTeach(path):
    print "Teaching classifier..."
    mapCounter = collections.Counter() #Hashmap equivalent
    files=glob.glob(path)
    for user in files:
        #print ' '
        f=open(user, 'r')
        for line in f:
            #lines at hand :)
            for word in clean(line): #Preprocessing here
                #words at hand :)
                mapCounter[word.lower()] +=1 #words should be converted
to lowercase
    return mapCounter
#TaskTeach end
```

CURRICULUM VITAE

Kaloma Usman Majikumna was born in 1989 at Damagum Fune LG. Yobe State Nigeria. Attended primary school at Damagum Central Primary School from 1995-2000 and secondary school at Government College Nguru from 2000 to 2006. Graduated with second class upper division honors in computer engineering from Melikşah University, Kayseri Turkey in 2014. He has been a Master Degree Student at Firat University Software Engineering Department since fall 2014.

CONTACT

Address: Abbari ward Damagum Fune LG. Yobe State Nigeria.

Email: kalomausman@mail.com