6-1-2014

Identifying the topic-specific influential users in Twitter

May Shalaby

THE AMERICAN UNIVERSITY IN CAIRO
SCHOOL OF SCIENCE AND ENGINEERING

Identifying the Topic-Specific Influential Users in Twitter

A Thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of Master of Science

By

May Shalaby

Under the Supervision of

Prof. Dr. Ahmed Rafea

Spring 2014

Abstract

Social Influence can be described as the ability to have an effect on the thoughts or actions of others. Influential members in online communities are becoming the new media to market products and sway opinions. Also, their guidance and recommendations can save some people the search time and assist their selective decision making.

The objective of this research is to detect the influential users in a specific topic on Twitter. In more detail, from a collection of tweets matching a specified query, we want to detect the influential users, in an online fashion. In order to address this objective, we first want to focus our search on the individuals who write in their personal accounts, so we investigate how we can differentiate between the personal and non-personal accounts. Secondly, we investigate which set of features can best lead us to the topic-specific influential users, and how these features can be expressed in a model to produce a ranked list of influential users. Finally, we look into the use of the language and if it can be used as a supporting feature for detecting the author's influence.

In order to decide on how to differentiate between the personal and non-personal accounts, we compared between the effectiveness of using SVM and using a manually assembled list of the non-personal accounts. In order to decide on the features that can best lead us to the influential users, we ran a few experiments on a set of features inspired from the literature. Two ranking methods were then developed, using feature combinations, to identify the candidate users for being influential. For evaluation we manually examined the users, looking at their tweets and profile page in order to decide on their influence. To address our final objective, we ran a few experiments to investigate if the SLM could be used to identify the influential users' tweets.

For user account classification into personal and non-personal accounts, the SVM was found to be domain independent, reliable and consistent with a precision of over 0.9. The results showed that over time the list performance deteriorates and when the domain of the test data was

changed, the SVM performed better than the list with higher precision and specificity values. We extracted eight independent features from a set of 12, and ran experiments on these eight and found that the best features at identifying influential users to be the Followers count, the Average Retweets count, The Average Retweets Frequency and the Age_Activity combination. Two ranking methods were developed and tested on a set of tweets retrieved using a specific query. In the first method, these best four features were combined in different ways. The best combination was the one that took the average of the Followers count and the Average Retweets count, producing a precision at 10 value of 0.9. In the second method, the users were ranked according to the eight independent features and the top 50 users of each were included in separate lists. The users were then ranked according to their appearance frequency in these lists. The best result was obtained when we considered the users who appeared in six or more of the lists, which resulted in a precision of 1.0. Both ranking methods were then conducted on 20 different collections of retrieved tweets to verify their effectiveness in detecting influential users, and to compare their performance. The best result was obtained by the second method, for the set of users who appeared in six or more of the lists, with the highest precision mean of 0.692. Finally, for the SLM, we found a correlation between the users' average Retweets counts and their tweets' perplexity values, which consolidates the hypothesis that SLM can be trained to detect the highly retweeted tweets. However, the use of the perplexity for identifying influential users resulted in very low precision values.

The contributions of this thesis can be summarized into the following. A method to classify the personal accounts was proposed. The features that help detecting influential users were identified to be the Followers count, the Average Retweets count, the Average Retweet Frequency and the Age_Activity combination. Two methods for identifying the influential users were proposed. Finally, the simplistic approach using SLM did not produce good results, and there is still a lot of work to be done for the SLM to be used for identifying influential users.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

The Internet is constantly developing into a highly interactive medium. During the last years, we have witnessed a massive transition in the applications and services hosted on the Web. The obsolete static Web sites have been replaced by novel, interactive services whose common feature is their dynamic content. The social and participatory characteristics that are included in these services has allowed the users to not only obtain information but also actively generate content, turning the former mass information consumers to the information producers. This has led to the generation of virtual communities, where users share their ideas, knowledge, experience, opinions and even media content (Akritidis et al., 2009). Examples include blogs, forums, wikis, media sharing, bookmarks sharing and many others, which are collectively known as the Web 2.0.

With the pervasive presence and ease of use of the Web, an increasing number of people with different backgrounds flock to the Web to conduct many previously inconceivable activities. One of the most important features introduced by the deployment of Web 2.0 is social networking, which now makes up a significant part of the Internet. The explosive growth of social media has provided millions of people the opportunity to create and share content on a scale barely imaginable a few years ago. Massive participation in these social networks is reflected in the countless number of opinions, news and product reviews that are constantly posted and discussed in social sites (Romero et al., 2010).Common examples are the popular social networking sites like Friendster, Facebook, MySpace, etc. Social media also includes YouTube, Photobucket, Flickr, and other sites aimed at photo and video sharing. News aggregation and online reference sources, examples of which are Digg and Wikipedia, are also

counted in the social media bucket. Blogging websites are also included, such as Blogspot, Wordpress and Tumblr, which allow people to update and share posts and links on their interests and everyday lives.

As the need for quick and short updates increases, micro-blogging is an emerging form of communication, especially with the furthering development of applications in the mobile domain, which means that users update from wherever they are and whenever they like. Micro-blogging allows users to publish brief message updates, which can be submitted in many different channels. One of the most notable micro-blogging services is Twitter, which allows users to publish posts, known as "tweets" with a limit of 140 characters (Weng et al., 2010).

Typically, users will tweet of topics that interest them. This may be related to their work, a hobby, or a mixture of multiple areas. These tweets are generally posted with the idea that they will be useful or interesting for some of the user's followers as well as an attempt to attract more followers. Twitter is also used as a means to contact friends and to get assistance and opinions on topics. Therefore, particular users may belong to different communities of people depending on what kind of posts they want to view (Webberley, 2011).The new self awareness of the information society has lead to the fact that more and more users connect online in social networks in order to exchange opinions. They interact with each other and influence each other's opinions (Bodendorf and Kaiser, 2009).

The notion of influence has long been studied in the fields of sociology, communication, marketing, and political science. Influence plays a vital role in how businesses operate and how society functions. In a real-world community, people tend to consult others when they are about to make a decision. Such decisions include purchases, event attendances, travel destinations, or even political voting. Similarly, online social networks are a virtual world, to which the users can connect to anywhere anytime. Users ask and listen to the opinions of fellow online users on

various aspects of life, such as which restaurant to choose, which place to visit, or which movie to watch. Hence, they are influenced by others in their decisions (Akritidis et al., 2011). That, electronic word-of-mouth, has become so important that the identification of the influential members can benefit all; in developing business opportunities, forging political agendas, discussing social and societal issues, and can also lead to many interesting innovative applications(Agrwal et al., 2008).

1.2 Motivation

Influential members and opinion leaders are usually well connected in large communities; consequently, they play a special role in multiple ways. The influential people in a society are often *market-movers*. Since they can affect buying decisions of their fellow users, identifying them can help companies better understand the key concerns and new trends about products interesting to them, and smartly respond to them with additional information and consultation to turn them into unofficial spokesmen. Apart from their commercial and advertising significance, influential members could also be responsible for forging political agendas by affecting the voting behavior of their readers; they could *sway* opinions in political campaigns, elections, and affect reactions to government policies (Drezner and Farrell, 2004)(Akritidis et al., 2011).

Word-of-mouth diffusion has long been regarded as an important mechanism by which information can reach large populations. It is believed that electronic word-of-mouth has greater influence than traditional marketing tools. The content created and consumed in online communities has become strategically important for companies and organizations interested in population feedback. Thus, tapping on the influential people in a community can help understand the changing interests, foresee potential pitfalls and likely gains, and adapt plans timely and pro-actively, not just reactively. Also, when faced with the massive amount of opinions generated, users can be overwhelmed and at loss about whose opinions are trustworthy. Influential members

can be helpful in giving recommendations, providing customer support and troubleshooting since their solutions are trustworthy because of the sense of authority these members possess (Agrwal et al., 2008).

Among social networking websites, one of the most important is Twitter. Twitter is offered in 21 languages, thanks to crowd-source translations from volunteer users (Twitter, 2012a). It has gained huge worldwide popularity since the first day that it was launched in 2006 and is now one of the most visited internet sites. Also, it is one of the fastest growing online social networking services, with a huge volume of content generated daily. Statistics show that halfway through 2011, users on Twitter were sending an average of 200 million Tweets per day. For context on the speed of Twitter's growth, in January of 2009, users sent two million Tweets a day, and in 2010 they posted 65 million a day (Twitter, 2011). As of March, 2012, there were over 140 million active users, and 340 million Tweets a day (Twitter, 2012b). Those numbers have, yet again, grown since then, and as of March 2013, there are well over 200 million active users creating over 400 million Tweets each day (Twitter, 2013).

Twitter has strongly influenced the way we communicate. For many people, it has become a part of their everyday lives. Twitter has made it easy to connect with friends and relatives, and to share thoughts, opinions and news with them from anywhere at any time. Twitter supports posting messages via SMS, web and mobile web services in addition to allowing users to use different third party applications to post and consume tweets. Also, it is extensively used in various fields as an easy, fast and convenient information sharing tool. The popularity of Twitter makes it an important tool for journalism, marketing, political campaigns and social change, and has thus drawn increasing interests from both the industry and research community.

Over the past 15 years, the Arab region has witnessed major technology-led transformations which changed the normal conduct in people's everyday life. The number of

individuals using the internet in the Arab region has reached 125 million, with more than 53 million actively using social networking technologies (Alshaer and Fadi, 2013).

Some interesting key findings by Alshaer and Fadi (2013) are, as of March 2013, the number of Twitter users in the Arab region has almost doubled in the last year to reach over 3.7 million users, with the highest number of active users, 1.9 million, in Saudi Arabia, which accounts for over half of all active Twitter users in the region. The estimated number of tweets produced by users in the region reached an average of 10 million tweets per day, with Arabic accounted for three quarters of the 336 million tweets sent in the region during March 2013. Saudi Arabia alone, produced almost half (47%) of all tweets in the Arab region, while Egypt produced 12% and the UAE produced 11%.

Twitter has been framed as an important news-bearing medium and has been touted for the role it played in the popular uprisings that have spread across the Arab region since December 2010. However, only 0.26% of the Egyptian population, 0.1% of the Tunisian population, and 0.04% of the Syrian population are active on Twitter. Of all the countries in North Africa and the Middle East, Twitter is most popular in Kuwait, where 8.6% of the population is active on Twitter (Fox, 2012). Nonetheless, as of the year 2013, tweets have become one of the most important sources of news in Egypt, as well as a tool for coordinating activism, events and protests. This rise in use of social media networks coincides with the explosive growth of smart phone use across the region in the past few years.

Webberley et al. (2011) described that the strength of Twitter is in its social structure and social networking functionality. Each user is allowed to choose who they want to follow; conversely, they may also be followed by others. Followers of a user receive all of that user's posts in their timelines. As a result, people are likely to follow users who update with interesting posts. If someone sees a post that they feel would be interesting to their followers, they can

'retweet' the post, rebroadcasting it to their own followers. The users whose updates are being followed by a particular user effectively become filters of information for that user. The user can choose to follow another user, and therefore implicitly indicates the kind of information they want to receive. A follower of a user may decide to retweet the retweeted tweet, thus creating a chain. Naturally, the larger the user's effective audience, both directly and through retweets, the greater the chance of being retweeted, again, and having their message spread. Suh et al. (2010) showed this by demonstrating how the retweet rate increases with the number of followers of the original tweeter. This is also related to the ideas of user influence mentioned by Cha et al. (2010). Bakshy et al. (2011) state that Twitter is well suited to studying influence, defining influencers as individuals who disproportionately impact the spread of information or some related behavior of interest. Ordinary individuals communicating with their friends may be considered influencers, but so may subject matter experts, journalists and other semi-public figures, as well as highly visible public figures like media representatives, celebrities and government officials. An especially useful feature of Twitter is that it not only encompasses various types of entities, but also forces them all to communicate in the same way, via tweets to their followers.

1.3 Research Objective

The objective of this research is to detect the topic-specific influential users on Twitter. From a collection of tweets matching a specified query, retrieved in reverse chronological order, we want to detect the relevant influential users, in an online fashion. In order to address our objective, a few questions were raised along the way.

First, we learn that not all Twitter accounts are the same. Since we want to focus our search for the influential users on the users who are personally involved, we investigate how we can differentiate between the personal and non-personal accounts.

Secondly, provided information about the tweets and their authors, we investigate which set of features can best lead us to the topic-specific influential users, and how these features can be expressed to produce a ranked list of influential users.

Finally, we look into the user's use of language in writing the tweets; if it can be used as an indicator of the author's influence.

1.4 Methodology

In order to decide on how to differentiate between the personal and non-personal accounts, we first retrieve from Twitter a collection of tweets for a number of topics. From the tweets we extract the unique users and their information. We determine the relevant user features for our problem and use them to carry out the account classification. We compare the effectiveness of account differentiation between the automated classification approach and a manually assembled list of non-personal accounts used as a comparative reference.

In order to determine the set of features to use in ranking the users and detecting the influential users, we first determine the set of relevant features. We retrieve a collection of tweets discussing a specific topic and extract the features. We then calculate the correlation values between each two features in order to gain a view of their dependencies.

We retrieve a collection of tweets discussing a specific topic. We study the collection users and prepare a manually assembled list of those we think are the influencers. Then using the independent features, we develop a number of ranking methods. To decide on the method that would best rank the users according to influence we evaluate outcome using the manually assembled list of influential users to calculate the precision. Finally, the influential users' precision values from the different ranking methods are tested for statistical significance.

To tackle our final investigation, we decided to use a statistical language model (SLM). The language model first needs to be trained with a considerably large corpus. So a large collection of tweets for a number of topics are retrieved and preprocessed to make up the language model training corpus. The language model is then developed and the hypothesis of whether it may be used to detect influential tweets is tested.

1.5 Thesis Layout

The rest of this document is organized as follows: Chapter 2 reviews the approaches covered in the literature for identifying influential members in online social networks. Chapter 3 describes the proposed approach, including the tools and methodologies used. Chapter 4 shows the experiments carried out for detecting influential users using the Twitter features and Chapter 5 shows the experiments carried out for using a Statistical Language Model for detecting influential users. Finally, in chapter 6, we conclude our work.

CHAPTER 2

Approaches for the Identification of Influential Members

Identifying influential users in online social networks, such as Twitter, has been actively studied recently. In this chapter, we briefly review some of the approaches studied for the identification of influential members in an online social network. The chapter is organized as follows. We first discuss the importance of influence, and then highlight some of the features used to define an influential member. After that, we lay out some of the approaches carried out for measuring users' influence, such as Social Network Analysis (SNA), the k-shell graph decomposition, a few mathematical models and algorithms, and also, the use of linguistic analysis. We then display a few of the evaluation approaches used in some of the studies in the literature, and finally finish up with our concluding remarks.

2.1 Importance of Influence

The Pareto principal (Pareto, 1971) exists almost everywhere. For example, 80% of a country's land is owned by 20% of the population, and 80% of a company's sales revenues comes from 20% of its clients. This is also the case for many social networks. In these networks, there exists the two types of users; those that exhibit different influence and different behavior. For instance, it has been shown by Wu et al. (2011) that less than 1% of the Twitter users (e.g. entertainers, politicians, writers) produce 50% percent of its content, while the others (e.g. fans, followers, readers) have much less influence and completely different social behavior (Weng et al., 2011).

Social Influence can be described as power; the ability of a person to have an effect on the thoughts or actions of others (Brown and Feng, 2011). Influential members and opinion leaders are becoming the media to transfer new products to consumers and sway opinions, whether deliberately or by chance. The guidance and recommendations of opinion leaders can save some

people the search time and assist their selective decision making more intuitively and comprehensively. Thus by reaching out to those opinion leaders, social media marketers willing to promote their campaigns may be able to trigger successful campaigns, and policy makers willing to promote social change may be able to promote real social change.

Automatically detecting influential members on online social networks has recently received great attention from both research and industry. However, the tools for measuring are still maturing and there is still no clear agreement over *what* to measure. Not all the tools measure the same kinds of things, so one may find several of these useful for their efforts and others not so much. Some may be useful for measuring a blog's or website's reach, while others assess popularity or presence on a particular social network.

2.2 How to define an Influential member

The search for influential members boils down to the question of how to define an influential member. First of all, active users are not necessarily influential and influential users can be inactive for periods of time. While active users can be simply defined by how frequently they publish new posts, it is a more complex matter how to define an influential user (Agarwal et al., 2008).

Following Keller and Berry (2003), one is influential if they are recognized by fellow citizens, can generate follow-up activities, have novel perspectives or ideas, and are often eloquent. Agarwal et al. (2008) set forth an initial set of intuitive properties that can be approximated by some collectable statistics. The same concept and some of the properties were also used by Akritidis et al. (2011).

• **Recognition** - An influential blog post is recognized by many. This can be equated to the case that an influential post is referenced in many other posts, the more influential the referring posts are, the more influential the referred post becomes. The concept is much like that in Web ranking

algorithms like PageRank and hyperlink-induced topic search (HITS), where links are used to convey authority.

• **Activity Generation** - A blog post's capability of generating activity can be indirectly measured by how many comments it receives the amount of discussion it initiates. In other words, few or no comment suggests little interest of fellow bloggers, thus non-influential. Hence, a large number of comments indicate that the post *affects* many such that they care to write comments and reply, and therefore, the post can be influential.

• **Novelty** - Novel ideas exert more influence as suggested in Keller and Berry (2003).

• **Eloquence** - An influential is often eloquent (Keller and Berry, 2003). There are many measures that quantify the goodness of a post such as fluency, rhetoric skills, vocabulary usage, and blog content analysis. This property is most difficult to approximate using some statistics. Given the informal nature of most social networks, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so. Therefore, the length of a post was used in Agarwal et al. (2008) and Akritidis et al. (2011) as a heuristic measure for checking if a post is influential or not. The blog post length was found to be positively correlated with number of comments, which means longer posts are likely to cause stronger reactions from the readers than shorter ones.

The above four form an initial set of properties possessed by an influential post. There are certainly other potential properties, such as user productivity and activity rate among many others. It is evident that each of the properties may not be sufficient on its own, and they should be used jointly in identifying influential bloggers.

In Agarwal et al. (2008), they broadly categorized the influential bloggers into the following *temporal patterns*: Long-term influential who steadily maintains the status of being influential for a very long time. They can be considered "authority" in the community. Average-term influential who maintains their influence status for a period of 4-5 months. Transient

influential who is influential for a very short time period (only one or two months).Burgeoning influential who is emerging as an influential blogger recently.

Disparate bloggers can present different temporal patterns. Long-term influential users are more influential than other bloggers as they are more trustworthy as compared to other bloggers based on a long time of history. Burgeoning influential users have potential to become long-term ones. But it is difficult to say these things about transient influential users as they might become influential by chance. Certainly, there could be many other temporal patterns depending on a particular application. On the other hand, regardless of temporal patterns, Akritidis et al. (2011) simply states that an Influential user would be recognized as such if they have several influential posts recently, or if the posts have had an impact recently. Findings by Zhou et al. (2009), however, indicate that in small communities, members' activities and the date they joined are of importance; the earlier they join and the more active they are, the more likely they will be considered as leaders.

In addition to user activity and connectivity, linguistic features have also been used to indentify influential members. Different types of individuals use language differently in their posts. Quercia et al. (2011) have studied one specific aspect that mediates interactions between users – their use of language – and have found that it is linked to social influence. They have found that language, with its vocabulary and prescribed ways of communicating, is a symbolic resource that they claim can be used on its own to influence others on Twitter; that influence partly depends on the linguistic qualities that reflect one's personality and mood.

There is another observation evident by the analysis presented in Agarwal et al. (2008) and Zhou et al. (2009), that many of the influential users are also active, i.e., productive. Although productivity and influence do not coincide, there is a strong relation between them (Akritidis et al., 2011).

2.3 Approaches for Measuring Users' Influence

In research, identifying influential users on online social networks such as Twitter has been actively studied recently. Much analysis on the data available has been done and there has been a broad spectrum of algorithms proposed.

Some might consider interpreting a Twitter user's influence as the number of followers they have. The more followers, the more impact the user has in the Twitter context. The underlying assumption is that every tweet published will be read by all the followers. However, this is not considered a good indicator of influence. In a dataset prepared for the study in (Weng et al., 2010), it was observed that 72.4% of the users follow more than 80% of their followers, and that 80.5% of the users have 80% of their friends follow them back. Reciprocity in the "following" relationships is prevalent in Twitter.

Weng et al. (2010) suggested two seemingly conflicting reasons that can possibly explain such reciprocity. First, the "following" relationship is so casual that each Twitter user just randomly follows someone, and those being followed follow back just for the sake of courtesy. Second, it might be the opposite; the "following" relationship is a strong indicator of the similarity among users. In other words, a Twitter user follows a user because they are interested in the topics that user publishes in tweets, and the user follows back because they find that they share similar topic interests. This phenomenon is called "homophily", which has been observed in many social networks (McPherson, 2001). If it is caused by the first reason, identifying the influential twitter user based on "following" relationship would be rendered meaningless since the following relationship itself does not carry strong indication of influence. On the other hand, the presence of homophily indicates that the "following" relationships between Twitter users are related to their topic similarity. Homophily is a phenomenon showing that people's social networks "are homogeneous with regard to many sociodemographic, behavioral, and

interpersonal characteristics" (McPherson, 2001).The presence of homophily implies that there are Twitter users who are serious in choosing friends to follow. This implication is important in that identifying the influential Twitter user based on the "following" relationships would be rendered meaningless if no twitter user is serious in "following" others (Weng et al., 2010).

In Twitter, empirical evidence supports that the idea that influencers are not accidental, but rather individuals who exhibit specific behaviors. Cha et al. (2010) describes influencers as individuals who keep great personal involvement and who limit their tweets to a single topic, and can thus be identified. Romero et al. (2010) found that influential individuals are highly-active users and consequently defined a new influence measure based on user activity. All this goes to show that influence on Twitter is not gained accidentally but strongly depends on audience engagement and user involvement (Quercia and Ellis, 2011).

Cha et al. (2010) compared three different measures of influence: in-degree, retweets, and mentions. Focusing on an individual's potential to lead others to engage in a certain act, Cha et al. (2010) highlighted three "interpersonal" activities on Twitter. Users interact by following updates of people who post interesting tweets. Users can pass along, by retweeting, interesting pieces of information to their followers. Finally, users can respond to, or comment on, other people's tweets, which is called mentioning. These three activities represent the different types of influence of a person. The number of followers of a user directly indicates the size of the audience for that user. The number of retweets indicates the ability of that user to generate content with pass-along value. The number of mentions containing one's name indicates the ability of that user to engage others in a conversation. The top users, based on each measure, showed a strong correlation in their tweet influence and mention influence. This means that users who get mentioned often also get retweeted often, and vice versa. The number of followers, however, was not related to the other measures. Cha et al. (2010) concluded that the most

connected users are not necessarily the most influential when it came to engaging one's audience in conversations and having one's message spread.

Cha el al. (2010) also found that the most influential users often hold significant influence over a variety of topics. This means that local opinion leaders and highly popular figures are trusted and could indeed be used to spread information outside their area of expertise. They found that influence is not gained spontaneously or accidentally, but through concerted effort. In order to gain and maintain influence, users need to keep great personal involvement.

2.4 Existing Tools for Social Network Analysis

Although Twitter hasn`t launched their Analytics tool yet, many others have made use of the Twitter API and built some apps. Companies like Klout and PeerIndex summarize social media activity and data, and calculate and assign a score which would reflect a user's social media capital, whether it be influence, engagement, reach or impact, or even all. For example, Klout measures several metrics such as reach, demand, engagement and velocity, in addition to a compound score combining them all. The upside of these services is that they are usually free, at least at some basic level, and some of them are able collect data from multiple social media profiles automatically. On Klout, for example one can connect not only Twitter and Facebook, but also YouTube, LinkedIn, foursquare, Instagram, Flickr, Tumblr, Blogger and Last.fm accounts to factor in to your score.

Several more tools which analyze Twitter presence:

- **EmpireAve**: It works like a stock index of social media users. The user's value on the index is both a factor of their social media activity and their use of Empire Avenue; not a true social score, but a means of comparison.
- **Formulists**: It provides several automated list functions to help the user fully categorize and track their growing follower base.

- **TurnRank**: It provides a score which is a reflection of how much attention a user's followers are directly giving and how much attention they bring you from their network followers.
- **TweetReach**: It allows a user to analyze how "far" their tweet has traveled based on their Tweet's total exposure. It also calculates the potential for their tweet to be seen, rather than just tracking a raw number of mentions and retweets.
- **SocialMention**: It can be set to track mentions of brands, keywords or hashtags while ranking top contributors. It can also allow a user to track mentions of their brand, for example, across Twitter, Facebook, blogs and forums in real time.

For an avid social media user, these tools are the easiest way to gauge their overall influence. They use almost real-time calculators to determine a user's overall impact, reach and value. They can also gain insight on a user's topics of influence, the users influenced the most, and the users who influence them. However, these social media measuring tools are not very definitive and not entirely reliable.

Clearly, there is a lot of room for research in social media evaluation. Measuring Influence, although challenging, is certainly not impossible, and there are several valid approaches.

In Twitter, several networks emerge from the user interactions enabled by the Twitter features. Various metrics and methods have been introduced for studying the "importance" of nodes within complex network structures. Such studies found applications in a variety of settings, the most prominent ones being the analysis of the Internet topology and the study of social networks.

2.4.1 Social Network Analysis

A social network is a set of people or groups of people with some patterns of interactions between them. They are useful for analyzing interactions that involve a large number of entities. Research on social networks could be traced back to sociology, anthropology and epidemiology. In due time, social scientists have developed it into a powerful tool: Social Network Analysis (SNA)

In theoretical frameworks, the formalization of social networks consists of the social graph and the social relationship matrix. A social graph consists of a number of nodes, each representing an entity, and their inter-links, which are usually directional links. A social graph can be expressed as a social relationship matrix which shows if there is a relationship between two nodes and its strength. Thus, in short, the nodes are the entities and the edges are the relationships

SNA is a quantitative social scientific method for measuring social relations through an emphasis of structural relations, which posits that the structure of social networks affects perceptions, beliefs, and actions through a variety of structural mechanisms that are socially constructed by relations among entities (Murthy et al., 2011).

Since SNA focuses on the interconnections of the actors, it is used to analyze the interpersonal relationships between various social actors within an organization or community and can provide rich and systematic descriptions and interpretations of complex social relationships. Researchers using SNA build behavioral models, describe to the best of their knowledge the structure of the intra-group relationships, and examine the influences of the network structure on the group as a whole and the behavior of the individual members.

This method can be used to describe and measure the relationship between network members and the flow of all kinds of tangible or intangible things through these relationships,

such as information, resources and knowledge (Ya-ting and Jing-min, 2011) and enables the examination of patterns in the relationships among interacting users. The goal is to identify opinion leaders and influential members, those people who play a crucial role in forming opinions and affecting others with their special position and communication habits within the network.

The literature on SNA is well established and so are the metrics and modes of visualization. There are many key figures in the field of social network analysis which describe the position and communication habits of users to analyze the user interaction network in order to find influential users. Sun and Qiu (2008) mentioned a few of the common concepts in Social Network Analysis:

1. **Degree**: The degree of a node is the number of links to this node. In a directed graph, there is an in-degree and an out-degree. The out degree of node is the number of links pointing out of this node, and the in-degree is the number of links pointing to the node. If both the in-degree and out-degree of node are zero, then the node is called an isolated node.
2. **Geodesic path**: There can be multiple paths of varying distance between any two given nodes. The shortest of all the paths between two nodes is called the geodesic path.
3. **Geodesic distance:** The distance of the geodesic path(s) between two nodes is called the geodesic distance, represented by $d(i,j)$. If no paths exist between two nodes, then the distance between them is infinite or undefined.
4. **Diameter:** A network graph generally has many geodesic paths with varying distance. The distance of the longest geodesic path is called the diameter D of the network, which may be formalized as $D = Max\{d(i,j)\}$.
5. **Density:** Density is a measure of the closeness of a network. Given a number of nodes n, the more links l between them, the larger the density. The density is $\rho = \frac{2l}{n*(n-1)}$ for directional graph and $\rho = \frac{l}{n*(n-1)}$ for undirected graph.

6. **Power and Centrality:** Power is an important concept in social network analysis. Social scientists measure power from the perspective of "relationship" and have given it many different formal definitions, including degree of centrality and centrality potential. Social network analysts tend to use "centrality" to express the concept of power. Centrality tells what central role a person or organization plays in a social network.

Through social network centrality analysis it is easy to find the core member in the network and relatively important members. The centrality analysis in social network analysis is mainly used to analyze the central position that an individual or organization is in its social network. Centricity index can be divided into two parts. One is the centrality of point and the other is the centrality of graph. The former usually describes the core locations of a single actor in network, and the latter describes the center trend of the network (Ya-ting and Jing-min, 2011).

The centrality of point usually can be divided into degree centrality, betweenness centrality and closeness centrality. Among them, the degree centrality measures actors' ability to interact; the betweenness centrality is used to measure actors' resource control ability; the closeness centrality describes the independence of actors from other actors. The calculation of betweenness centrality and closeness centrality depends on the relationship between one and all the other actors in the network, not just the direct relation between neighbors (Ya-ting and Jing-min, 2011).

Degree Centrality Analysis is a measure of direct connections. Generally speaking, if a member has direct association with many other members, then the member is in central position. Under the guidance of this kind of thinking, the calculation of one point's degree centrality can use the number of points which have a direct relationship with the point. In other words, degree centrality of a point is the comprehensive of the out-degree and in-degree. However, this measure can be misleading, since just increasing a member's direct connections won't increase their influence.

Betweenness Centrality Analysis is measures how well positioned a member is. If a member is in the shortest path between many other actors, this actor is in an important position. According to this kind of thinking, betweenness centrality can be used to measure the resources control degree of the actor.

Closeness Centrality is measure is somewhat a synthesis of the previous two. It measures the amount of social distance a node would travel to get to anyone else in the network using both direct and indirect links. It is known that if one member is less dependent on others in the contacting process, they have higher centrality. According to closeness centricity (namely a point is much closer with other points, its communication with the outside world is more independent), closeness centrality index can depict the center index. It is in the important bridging position in the network, and plays an important role in network transmission. The interactions of many other members often depend on it.

2.4.2 Recent Studies using SNA to identify Influential members

After having obtained the link relationships between the members of the blogosphere, Sun and Qiu (2008) used Social Network Analysis to explore the structural features of the blogosphere and the behavioral patterns of its members. They focused, in their analysis, on the degree centrality of a node; this is measured as the sum total of the in- and out-degree of the node, which reflects the strength of attention paid to this node by others.

Bodendorf and Kaiser (2009) used Social Network Analysis to detect opinion leaders and opinion trends. They proposed a new approach which detects opinions and relationships among forum users by text mining. On this basis, the main influential factors for opinion forming in virtual communities are extracted. By social network analysis metrics, opinion leaders were identified and opinion evolvement is analyzed.

Cui et al. (2011) adopts social network theory to study the topology characteristics and features of blogosphere. They used key performance indicators including degree distribution, both in-degree and out-degree, average Geodesic distance length, clustering coefficient, which is a measure of degree to which nodes in a network tend to cluster together, and spectral density, which captures the frequency content of a stochastic process and helps identify periodicities.

Bigonha and Cardoso (2010) proposed a for ranking the most influential users on Twitter based on a combination of the user position in the network topology, the polarity of that user's opinions and the textual quality of the tweets. They defined the influence of a user based on their network position and their behavior – the interaction with other users, the polarity of the user's opinions and the quality of the posted tweets. Given a certain topic, they defined evangelists and detractors, the influential users who act in favor and against a subject, respectively.

From the several networks that naturally emerge from the user interactions enabled by the Twitter features, Bigonha and Cardoso (2010) selected two of them for an in-depth analysis: Follower/Following Network and Interactions Network. The most common interactions are replies, in which one user wants to answer a post from another, and retweets, directing a post from another to that user's followers. They used a number of graph network metrics to analyze the user interaction network, specifically at individual node properties, such as degree, betweenness and centrality. The graph network metrics were used in combination with the TFF Ratio (Twitter Follower-Friend Ratio): the ratio of a user's followers to friends (people who the user follows), to identify influential users in a dataset, considering the users with higher TFF Ratio as more relevant. Since the number of profiles following a user in not directly related to influence, but is an indication of that user's popularity (Cha et al., 2010). As for measuring how well written and understandable a tweet is, Bigonha and Cardoso (2010) used the Flesch-Kinkaid metric (Graber et al., 1999). Their experimental results demonstrated that the s they used were

successful in identifying some of the most influential users, and that the interactions between users are the best evidence to determine user influence.

Murthy et al. (2011) used Social Network Analysis to understand complex health networks in social media, which were described as fluid, resist traditional notions of trust, and often lack explicit bidirectional relationships. Their goal was to develop an approach fusing social network analysis, natural language processing, and machine learning to analyze confirmatory and negatory mentions in social media and how they were being responded to in order to determine the flow of health information, trust, resources and ideas on social media and their impact on health outcomes. Also, the authority of individual Twitter users was analyzed with using social network analysis. This was used to better understand why users would trust particular health messages and what impact these relationships have on bettering health outcomes.

Ya-ting and Jing-min (2011) used Social Network Centrality Analysis methods to analyze a political blog community in order to find the core group members, the relatively important members and members with special characteristics. They used social network analysis to describe and measure the relationship between network members and the flow of all kinds of tangible and intangible things through these relationships, such as information, resources and knowledge. The centrality analysis in social network analysis is mainly used to analyze the central position that an individual or organization is in the network. Centricity index can be divided into two parts. One is the centrality of point and the other is the centrality of graph. The former usually describes the core locations of a single actor in network, and the latter describes the center trend of the network.

2.5 The k-shell graph decomposition algorithm

The k-shell decomposition algorithm is a well established method for detecting the core and the hierarchical structure of a given network. It has founded a number of applications as a

means for understanding the "importance" of nodes within large–scale network structures (Carmi et al., 2006).

Viewed as nodes in a graph, the higher the k-shell level assigned, the closer the node is to the core of the graph. Kitsak et al. (2010) proposed the use of K-shell decomposition as a technique for identifying the most influential spreaders in a complex network. The assumption is that if the nodes of the graph are users in a social network, the users in the high k-shell levels are more influential than in the network than users in lower k-shell levels.

The k-shell decomposition algorithm groups all nodes in a network that have k, or less, connections or that are only connected to other nodes with k, or less, connections. Once a node has been identified, it is marked and the search continues until all nodes in the k-shell have been found. The process then moves on to the next larger k-shell value, and continue until all nodes have been marked. In this basic algorithm, the k-shell values are assigned in a linear fashion; each k-shell value is equivalent to the analyzed connection count.

The algorithm is simple in theory, however, in practice it can be very time consuming, especially for a large network such as Twitter. Brown and Feng (2011) investigated a modified k-shell decomposition algorithm for computing user influence on Twitter.

Initially they used the basic algorithm, as described, but the results turned out to be highly skewed with most of the users falling into the first few, low, k-shell levels, and the remaining users tailing off over thousands of additional higher k-shells. This distribution made statistical observation hard, and thus they modified the original algorithm by applying a logarithmic mapping.

In the modified algorithm, each k-shell level represents roughly the log value of the analyzed connection count, and so it places nodes with $2^k - 1$, or less, connections into k-shell level k, effectively consolidating the higher k-shell levels. This modified algorithm produces fewer and

more meaningful k-shell values and a more useful distribution, in addition to being faster than the original.

2.6 Mathematical Models and Algorithms

Agarwal et al. (2008) was one of the first to propose a model attempting to quantify an influential blogger. They suggested that an intuitive way of defining an influential blogger is to check if the blogger has any influential posts, i.e., *A blogger can be influential if they have more than one influential blog post.* They proposed a preliminary model, using an initial set of intuitive properties supposedly possessed by an influential post, which allows for evaluating different key measures for identifying the influential members. Also, by tuning the weights associated with the parameters, the model can be adapted to look for different types of influential bloggers, and be used to examine how the different parameters impact the influence ranking.

The set of properties; *recognition* (ι) which is reflected in the number of in-links referencing post p; *activity generation* (γ) which is reflected in the number of comments the post received ; *novelty* (θ) which is reflected in the number of out-links and *eloquence* (λ) which is reflected in the length of the post. The properties used jointly to calculate an influence score $I(p_i)$ for the post p_i, which is determined by the following equation:

$$I(p) = w(\lambda)(\, w_{com}\gamma_p + w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n))$$

where $w(\lambda)$ is the weight function depending on the length λ of the post, w_{com}, w_{in} and w_{out} are the weights used to adjust the contribution of comments, ingoing and outgoing influence respectively. So for a blogger b_k who has *N* blog posts, their influence scores can be ranked in descending order, and the influence index of the blogger, $iIndex(b_k)$ can be defined as $max(I(p_i))$. Thus the problem of identifying the influential bloggers is defined as

determining an ordered subset of K that are ordered according to their $iIndex$ (Agarwal et al., 2008).

However, Akritidis et al. (2009) argued that isolating a single post to identify whether a blogger is influential or not is an over simplistic approach. They think that the productivity of a blogger is a significant issue that has been overlooked by the model of (Agarwal et al., 2008). Although productivity and influence do not coincide, there is quite a strong relationship between them, and therefore should somehow be taken into account. Also, they argue that the outcome of the model is not objective, since it depends highly on user defined weights, and most importantly, the model ignored what they considered to be one of the most important factors: The temporal dimension. Virtual social networks are rapidly changing environments, in a manner that a blogger who would currently be considered as an influential, is not guaranteed to remain influential in the future; an issue being discussed in a post at the present time and is now of major importance, may be totally outdated after a couple of months, or even days. An effective model should take into consideration the age of a post and also the age of incoming links to that post, in order to be able to identify the *now influencers*. Motivated by these observations, Akritidis et al. (2009) propose two easily computed blogger ranking methods, which incorporate temporal aspects of the blogging activity. The first metric, termed MEIBI (*Metric for Evaluating and Identifying a Blogger's Influence*) takes into consideration the number of the blog post's incoming links and its comments, along with the publication date of the post. On the other hand, an old post may still be influential. This could be deduced by examining the age of the incoming links to this post. And so, the second metric, MEIBIX (*MEIBI eXtended*), is used to score a blog post according to the number and age of the blog post's incoming links and its comments.

For both the MEIBI and MEIBIX no user defined weights need to be set to provide results, whereas the most sound features of blogs are considered. To an extent MEIBI and MEIBIX

produce similar rankings, however, MEIBIX is more affected by the number of incoming links, whereas MEIBI assigns better scores to the posts that attracted more comments.

Akritidis et al. (2011) then investigated the issue of identifying bloggers who are both productive and influential by introducing the blogger's productivity index and blogger's influence index. They identified a few factors that play a crucial role in the measurement of a blogger's influence and proposed two time-aware metrics. For the metrics proposed, they considered both the temporal and productivity aspects of the blogger's behavior, along with the inter-linkage among the posts. The first metric, *blogger's productivity* (BP) index, is used to evaluate the productivity of a blogger with respect to recency. So a blogger is considered to be currently productive if they have posted several long posts recently. The second metric, *blogger's influence* (BI) index reflects the influence of a blogger inside and outside a community by taking into consideration the number and age of the incoming links and comments. So for identifying influential bloggers, they are bloggers whose posts are receiving many comments and incoming links presently. The combination of these two values, BI-Index and BP-Index, can be used to characterize the bloggers.

Romero et al. (2010) added that it is important to also take into consideration the passivity of members of the network. The passivity of some users provides a barrier to the information propagation, which is often difficult to overcome. Romero et al. (2010) proposed an algorithm that determines the influence and passivity of users based on their information forwarding activity. The passivity of a user is a measure of how difficult it is for others to influence him, and the influence of a user depends on both the quantity and the quality of the audience, or followers and friends. The proposed model makes the following assumptions:

1. A user's influence score depends on the number of people they influence as well as their passivity.

2. A user's influence score depends on how dedicated the people they influences are. Dedication is measured by the amount of attention a user pays to a given one as compared to everyone else.
3. A user's passivity score depends on the influence of those who they're exposed to but not influenced by.
4. A user's passivity score depends on how much they reject other user's influence compared to everyone else.

The algorithm iteratively computes both the passivity and influence scores simultaneously. The IP algorithm outputs a function $I: N \rightarrow [0;\ 1]$, which represents the nodes' relative influence on the network, and a function $P: N \rightarrow [0;\ 1]$ which represents the nodes' relative passivity of the network (Romero et al., 2010).

Zhou et al. (2009) introduced the concept of Opinion Networks, and proposed a PageRank-like algorithm, to rank nodes in an opinion network. An opinion network is a directed graph with a set of nodes, each representing a member of the community or a group, and its edge set, where each edge represents an opinion orientation, and also a set of opinion scores associated with the edges.

Weng et al. (2010) proposed TwitterRank, an extension of PageRank algorithm, to measure the influence of users in Twitter. Firstly, the use of PageRank is motivated by the idea that the influence of a Twitter user can be interpreted similar to the "authority" of a webpage; a Twitter user has high influence if the influence of their followers is high, at the same time, their influence on each follower is determined by the relative amount of content the follower receives from them. Secondly, since the influence of a Twitter user may vary in different topics, the topic-sensitive algorithm, TwitterRank, was proposed to measure a user's influence. TwitterRank measures the influence taking both the topical similarity between users and the link structure into account.

Forming a directed graph of the "following" relationships among the users, a random surfer visits each user with a certain probability by following the appropriate edge. TwitterRank differentiates itself from PageRank in that the random surfer performs a topic-specific random walk, where the transition probability from one user to another is topic-specific. The more similar the two users are; the probability that the two users are interested in the same topic, the higher the transition probability from one to another. By doing so, topic-specific relationships are constructed among the users in the graph.

Bakshy et al. (2011) investigated the attributes and relative influence of Twitter users by tracking 74 million diffusion events that took place over a two month interval in 2009. Their use of the term influencer corresponds to a particular and somewhat narrow definition, specifically the user's ability to post URLs which diffuse through the Twitter follower graph, restricting the study to users who seed URL content. They measure influence in terms of the size of the entire diffusion tree associated with each event, as the size of the diffusion tree is directly associated with diffusion and the dissemination of information. So to calculate the influence score for a given URL post, they track the diffusion of the URL from its origin at a particular seed node through a series of reposts by that user's followers, those users' followers, and so on, until the diffusion event is terminated. They stated three choices for how to assign the corresponding influence: first, full credit is assigned to the friend who posted it first, rewarding primacy; second, full credit is also assigned to the friend who posted it last, attributing influence to the most recent exposure ; and third, credit is split equally among all prior posting friends, assuming that the likelihood of noticing a new piece of information, and the inclination to act on it, accumulates steadily as the information is posted by more friends. Disjoint influence trees are then constructed for every initial posting of a URL. The number of users in these trees defines the influence score for each seed. So for each user they aggregate all URL posts and compute the individual-level influence as a logarithm of the average side of all the influence trees for which that user was a

seed. They then fit a regression tree model (Breiman et al., 1984), in which a greedy optimization process recursively partitions the feature space, resulting in a piecewise-constant function where the value in each partition is fit to the mean of the corresponding training data.

They observed that the largest influence trees tend to be generated by users who have been influential in the past and who have a large number of followers. They found that the nature of the content did not necessarily improve predictive performance, but that individual-level attributes, in particular past local influence and the number of followers, can be used to predict average future influence.

2.7 Linguistic Analysis

While the style of writing used on Twitter is widely varied, much of the text is similar to SMS text messages. This is likely because many users access Twitter through mobile devices. Posts are often ungrammatical and filled with spelling errors. Twitter's noisy style makes processing the text more difficult than other domains. Nonetheless, the textual content has taken the interest of some studies. We have observed the use of statistical natural language processing s applied to the micro-blogging content.

Linguistic style has been central to a series of natural language processing applications, like authorship attribution, forensic linguistics, gender detection and personality type detection. Linguistic style is known to be generated and processed unconsciously. It is where style denotes the components of the language that are unrelated to content: how things are said as opposed to what is said (Danescu-Niculescu-Mizil et al., 2011). This is a rather important dimension, since, even though only 0.05% of the English vocabulary is composed of style words, an estimated 55% of all words people employ are style words (Tausczik and Pennebaker, 2010)

Kiciman (2010) examines the extent to which differences in language models in Twitter posts were related to the metadata associated with the senders, demonstrating the importance of

linguistic style variations in Twitter. While the textual content itself is quite short, there is a rich meta-data associated with every post, such as name, location and social details of the user; and easily inferred content meta-data, such as whether the post is a retweet, a reply, contains a web link, or whether other users or topics are explicitly referred. The hypothesis is that if a strong relationship exists between metadata features and language, then this meta-data can be used as a trivial classifier to match individual messages with specialized, more accurate language models.

A sample of 72 million Twitter posts was collected, preprocessed, and the English portion of the corpus was divided into subsets based on feature values. Separate n-grams language models were then learned for each of the subsets. For each feature studied, language differences were quantified by measuring the perplexity of each of the learned n-gram models against each subset of data. The results show that some metadata is correlated with language style, for example, the correlation between geography, provided by the "time-zone" as geographic location indicator, and language style. It is natural to expect that geography have an impact on language style due to language dialects as well as geographic-specific topics, events, place names, etc. Also, there was noticeable difference in the language among the groupings of posts whose authors had less than 10, 100 and 1000 followers. The largest language difference occurring among posts whose authors had more than 1000 followers. The main difference was in the use of ego-centric words, such as 'I', 'me' or 'my', as well as in words that are indicative of how one uses Twitter, for example, words like 'RT' indicating retweeting, and URL referencing in the post. While the use of ego-centric words doesn't vary significantly for user groups with less than 1000 followers, there is a significant drop in the use of these words by users with more than 1000 followers. Users with different numbers of followers also appear to retweet and reference web pages at different frequencies. The 'RT' token is likely to appear with users with fewer followers, and URL referencing is more probable from user groups with either less than 10 or more than 1000 followers.

Baron et al. (2012) investigates identifying social behavior; the presence of adversarial behavior and influence, of participants in online discussion forums from their language use, in English, Arabic and Chinese. The system they built uses a variety of features to predict the presence of social constructs. Given a thread of conversation, the system predicts the most salient posters that exhibit a target social construct in three phases. The first is message level processing in which each message in the conversation is analyzed, using a support vector machine (SVM) and linguistic evidence is collected. The SVM uses a variety of linguistic features from the message to make predictions. Second, that evidence is aggregated for each poster and used in poster level processing to decide if the poster exhibits the social construct and estimate confidence in each prediction. The confidence level output from the system range [0, 1] which is calculated from raw activation output by an SVM with sigmoid-like function. Third, this information is used to pick the most salient posters for the conversation using the confidence scores as a reflection of how much the system believes the behavior is present.

The effectiveness of the features used by the poster-level classifier was analyzed, grouping related features together to reflect the hypothesized social intent the feature is capturing. For adversarial behavior, there were commonalities across English, Arabic and Chinese; while with influence, the features that contribute positively to prediction differ across the language.

2.8 Evaluation Approaches

There seems to be no training and testing data to evaluate the efficiency of a proposed approach. The absence of ground truth about influential bloggers presents another challenge. The key issue is how to find a reasonable reference point.

As an alternative to the ground truth, Agarwal et al. (2008), one of the first to study influential bloggers, resorted to another Web2.0 site Digg (http://www.digg.com) to provide a reference point. According to Digg, “Digg is all about user powered content. Everything is

submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that's important to you!"

As people read articles or blog posts, they can give their votes in the form of dig and these votes are recorded on Digg servers. This means, blog posts that appear on Digg are liked by their readers. The higher the Digg score for a blog post is, the more it is liked. In a way, Digg can be considered as a large online user survey. Though only submitted blog posts are voted, Digg offers a way for us to evaluate the blog posts. Given the nature of Digg, a not-liked blog post will not be submitted thus will not appear in Digg.

As the Digg API only returns the top 100 voted posts, they use these 100 blog posts at Digg as the benchmark in evaluation. They would rank the blog posts based on their influence score and pick the top posts to be compared with the Digg set of 100 blogs to see how many also appear in the Digg set.

Akritidis et al. (2009) compared the influential bloggers indicated by their proposed methods to the bloggers found by H-index (Wikipedia, 2012) and those found by the influence-flow method proposed by Agarwal et al. (2008), both as state-of-the-art influential blogger identification methods. In addition to the state-of-the-art methods used by Akritidis et al. (2009), Akritidis et al. (2011) evaluated their proposed methods against the methods reported in Akritidis et al. (2009).

Romero et al. (2010) and Bakshy et al. (2011) both resorted to Bit.ly (https://bitly.com). They carried out their influence measure experimentations on tweets that included bit.ly URLs. Bit.ly is a URL shortening service that for each shortened URL keeps track of how many times it has been accessed, so the bit.ly URLs found in tweets can be queried for the number of clicks the service has registered on that URL. The URL click data was used to test how well the influence measure can predict the attention the URLs posted by the users receive. However, there is a wide

range of factors that can affect the click data which may affect the prediction accuracy. The main reason for that is that is that the amount of attention a URL gets is not only a function of the influence of users mentioning it, but also of other factors, including the virality of the URL itself and whether the URL was mentioned elsewhere (Romero et al., 2010).

Zhou et al. (2009) constructed a *Golden Standard* from a *real trust network* collected from Epinions (http://www.epinions.com), an e-commerce site where users can declare a list of members whom they trust. Based on the declared *trust list*, the nodes were ranked in the *real trust network* according to their in-degrees. They would then use *KSim* (Haveliwala, 2002) to measure the similarity between each of their methods' ranking results and the *Golden Standard*.

Bigonha and Cardoso (2010) did something quite similar, but a bit more labor intensive; they got a marketing and communications specialist to create a list of influential users for the studied theme. Among the users in the dataset, the specialist identified 17 influential users. So assuming the specialist's list as a ground truth, the proposed technique was assessed using several performance measures; precision, recall, average precision and mean average precision (Baeza-Yates and Ribeiro-Neto, 1999).

Another measure mentioned on several occasions in the literature is Klout. Klout is an internet service that claims to measure an individual's influence by aggregating information from a variety of social media platforms. Anger and Kittl (2011) mentioned Klout as one of the existing online rating services which determines user performance on Twitter, Facebook and LinkedIn. Klout measures, as stated on its website (http://www.klout.com), a user's overall online influence with a score ranging from 1 to 100, with higher scores corresponding to a higher assessment by Klout of the breadth and strength of one's online influence. It measures the size of a person's network, the content created, and how other people interact with that content. Klout analyses more than 25 variables, and offers to combine scores from all analyzed platforms. The

exact algorithm used to calculate the score is not published but Klout states that it sees influence as the "ability to drive people into action", thus making replies and retweets the most important factors.

Purohit et al. (2011) included each author's Klout score in their list of author features to study. Vega et al. (2010) randomly selected users attending a specific conference, on whose tweets the analysis was to be performed, based on their Klout influence score. They selected 20 users whose scores ranged between 24 and 84 out of the possible 100 points. This distribution was made to ensure that the users picked for the study had varied influence levels among other Twitter users.

Also, Klout was mentioned by Campo-Ávila et al. (2011) as one of the analytic tools used to calculate influence to obtain and compare data. However one of their remarks mention that they were unable to induce any accurate model for Klout, since the relations between the parameters are not as direct as some of the other tools. Even though it is known that 25 or more parameters are used to calculate the influence, none are provided. They concluded that some current tools may help measure how influential a twitter user is, but none provide an accurate measure of a standardized reach or scope by themselves.

Other than that, the evaluation of the different influence measures is usually done manually, like in Weng et al. (2010) and Ya-ting and Jing-min (2011) among many others. They would refer back to those ranked as high or low and observe their posts' contents and frequency, interaction with other members, maybe even their activity history, and return with some statistics which the model can be evaluated against.

It is obvious that the creditability of each evaluation approach highly relies on the type of data being analyzed, or vice versa; where some studies would customize the dataset and scope of the research to be able to use a certain reference point for evaluation.

2.9 Concluding Remarks

The number of followers may seem like an obvious and straight forward indication to a person being influential. It is in fact an over simplistic approach that gives indication of popularity and not necessarily that that person is influential. As mentioned by Cha et al. (2010) and proved by a few preliminary experiments we carried out, most highly followed users span a wide variety of public figures and news sources, showing that the most connected users are not necessarily the most influential. Even thought the number of followers can give an indication of the size of a user's audience, it may actually be inaccurate. Besides the fact that not all followers of a user read every tweet they posted, people don't necessarily need to be following a certain user to read their tweets, since Twitter users often use the search functionality to read the tweets mentioning or discussing a topic of current interest. Also there is the possibility that a percentage of a users followers be made up of inactive accounts if not fake spam accounts. So the number of followers may be taken into consideration as a contributing factor to a person's influence strength; the more the followers, the more the message is likely to spread, but it is not very reliable and cannot be used on its own.

On another note, Twitter being a network, analyzing the Twitter network using Social Network Analysis metrics may seem like the obvious way to go. However, due to the many features enabled by Twitter, several networks emerge, and not all of the relations can be integrated to a single network. That would require massive amounts of data collection for multiple network reconstructions. The constructed networks will only be snapshots of an instant in the life of the highly transitory Twitter network. Offline periodic analysis, with incremental updates is possible, but the speed at which content evolves makes it technically challenging, especially that some feature are very dynamic and liable to change faster than others; almost on an hourly basis. Not only is the size of the network affected, but its internal structure is highly

subject to change. So the application of network analysis on Twitter is highly impractical for real-time analysis of the dynamic rapidly ever-changing state of the network.

Following the research directed at analyzing the users and network and developing models to give bloggers and/or their posts scores based on how strong their impact or influence is, we recognize that influential members are usually individuals who exhibit specific behaviors. From the literature we can take away a few conclusions that will help in the founding of our model.

- An influential post will be recognized by many.
- The most connected users, even though they have a better chance at having their message spread more, they are not necessarily the most influential.
- Influence is not gained spontaneously or accidentally. Influential users often exhibit a few qualities, such as personal involvement, consistent activity rate and productivity. It also depends on audience engagement.
- Most influential users often hold significant influence over a variety of topics, however, the influence strength is bound to vary across topic genres.

CHAPTER 3

The Proposed Approach

With the rise in popularity and size of social media, there is a growing need for systems that can extract useful information from this amount of data. The micro-blogging service Twitter has evolved into a very popular tool for expressing opinions, broadcasting news, and simply communicating with friends. People often comment on events in real time, with several hundred micro-blogs (*tweets*) posted each second for significant events. Twitter is not only interesting because of this real-time response, but also because it is sometimes ahead of the newswire, with users posting eyewitness news. Among the millions of users, a small percentage is what is called the group of influencers. We address the problem of detecting the influential micro-bloggers using Twitter, taking into consideration that a user's influence may vary by topic genres. Another attractive feature to Twitter is the ease of use of its API to retrieve the necessary data to study.

Twitter basically being a network, it may be obvious that analyzing the network using Social Network Analysis and/or K-Shell graph decomposition algorithms seems like the obvious approach. However, studying the network requires that we collect friend/follower information and interaction information, such as retweets, mentions and replies, for network reconstruction. We were able to collect the necessary data, but to a very limited scale due to the API limitations. Huge amounts of data are required for network reconstruction. However, we were faced with the following issues. Firstly, the collectable data is not enough to capture a representation of a sub-network or do it justice in size and complexity. Secondly, the data collection process is very time consuming, due to the substantial amounts of data and metadata requested, in addition to the API rate limitations. Also due to the many features enabled by Twitter, several networks emerge, and not all of the relations can be integrated to a single network. And last but not least, the network constructed in the end is just a snapshot of an instant in the life of the network.

An offline periodic analysis, with incremental updates, would have been possible for the collection of a substantial amount of data, use the data to reconstruct the network and apply the SNA methods and/or the K-Shell graph decomposition algorithm on a snapshot of the network that week or day. However the Twitter network is highly transitory. When news breaks on Twitter, whether local or global, of narrow or broad interest, Twitter users flock to the service to find out what's happening. The speed at which content evolves makes it more technically challenging. The most frequent terms in one hour or day tend to be very different from those in the next, significantly more so on Twitter than in other content on the web. 17% of the top 1000 query terms "churn over" on an hourly basis. Repeating this at a granularity of days instead of hours, 13% of the top 1000 query terms from one day are no longer in the top the following day. During major events, the frequency of queries spikes dramatically (Twitter, 2012c). This rapid change alters the network just as fast. Not only is the size of the network affected, but its internal structure is subject to change. So the application of network analysis on Twitter is highly impractical for real-time analysis of the dynamic rapidly ever-changing state of the network.

For detecting influential members on Twitter discussing a certain topic, we propose the following approach. The first and most important step is to develop a data collection tool to retrieve the necessary data from Twitter on which our analysis is to be carried out. We then filter out the non-personal accounts in order to focus on the personal account users. A number of influential user ranking s are then developed and evaluated. Finally, the use of a statistical language model for tweet text evaluation is investigated to see if the user's language may be used as an influence indicator. The rest of this chapter explains how we approached each of these steps in detail.

3.1 Collecting Data from Twitter

Twitter is an information network and communication mechanism that produces more than 400 million tweets a day. So in order to identify the influential members tweeting about certain topics, we will need to retrieve and analyze the data available about the active users on Twitter. Luckily, the Twitter platform offers access to that corpus of data, via APIs. Twitter has two APIs. The Twitter REST API methods allow developers to access core Twitter data. This includes updating timelines, status data, and user information. It also includes the Search methods which allow developers to retrieve Twitter Search data. The Streaming API provides near real-time high-volume access to Tweets in sampled and filtered form. The Streaming API is distinct from the REST API as Streaming supports long-lived connections on a different architecture.

A tweets retrieval tool was developed making use of the Twitter REST API v1.1. It returns a collection of relevant Tweets matching a specified query, accompanied by some relevant metadata; user and tweet information. Multiple queries, using different search keywords, were retrieved. However, for each of our investigations there are different data requirements; each of the investigations uses a different set of the accompanying features and the amount of data used also differs.

Each tweet retrieved from Twitter via the API is accompanied by the following features viewed in Table 1, which lists each of the features and a brief description (Twitter API documentation, 2013a) (Twitter API documentation, 2013b).

Productivity and personal involvement are characteristics of influential users, who are known to voice their opinions and often take the initiative, those who generate the content others read. So, using the *retweeted_status* field, shown in Table 1, we are able to retrieve the original tweets and their data. When a retweeted tweet is encountered in the query result, the original tweet and its information is captured and the retweet is archived and associated with the original

tweet. So the queried tweets collection is made up of original content, and the noise and repetition caused by retweets are minimized.

Table 1: The list of features, retrieved from Twitter via the API, accompanying each tweet

Field		Description
created_at		UTC time when this tweet was created.
id		The integer representation of the unique identifier for this Tweet.
id_str		The string representation of the unique identifier for this Tweet.
text		The actual UTF-8 text of the status update.
entities		Entities which have been parsed out of the text of the tweet, such as the urls, hashtags and user_mentions.
retweet_count		Number of times this Tweet has been retweeted.
favorite_count		Indicates approximately how many times this Tweet has been "favorited" by Twitter users.
retweeted_status		Retweets can be distinguished from typical Tweets by the existence of a retweeted_status attribute. This attribute contains a representation of the *original* Tweet that was retweeted. Note that retweets of retweets do not show representations of the intermediary retweet, but only the original tweet.
in_reply_to_screen_name		*Nullable.* If the represented Tweet is a reply, this field will contain the screen name of the original Tweet's author.
in_reply_to_status_id		*Nullable.* If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID.
in_reply_to_status_id_str		*Nullable.* If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's ID.
in_reply_to_user_id		*Nullable.* If the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet.
in_reply_to_user_id_str		*Nullable.* If the represented Tweet is a reply, this field will contain the string representation of the original Tweet's author ID. This will not necessarily always be the user directly mentioned in the Tweet.
lang		*Nullable*. When present, indicates a BCP 47 language identifier corresponding to the machine-detected language of the Tweet text, or "und" if no language could be detected.
coordinates		*Nullable.* Represents the geographic location of this tweet as reported by the user or client application.
place		*Nullable.* When present, indicates that the tweet is associated (but not necessarily originating from) a Place.
contributors		*(Collection of Nullable)* A collection of brief user objects (usually only one) indicating users who contribute to the authorship of the tweet, on behalf of the official tweet author.
User	name	The name of the user, as they've defined it. Not necessarily a person's name.
	screen_name	The screen name, handle, or alias that this user identifies themselves with. screen_names are unique but subject to change

Field		Description
	created_at	The UTC datetime that the user account was created on Twitter.
	id	The integer representation of the unique identifier for this User.
	id_str	The string representation of the unique identifier for this User.
	protected	When true, indicates that this user has chosen to protect their Tweets.
	statuses_count	The number of tweets (including retweets) issued by the user.
	followers_count	The number of followers this account currently has.
	friends_count	The number of users this account is following (AKA their "followings").
	listed_count	The number of public lists that this user is a member of.
	favorites_count	The number of tweets this user has favorited in the account's lifetime.
	description	*Nullable*. The user-defined UTF-8 string describing their account.
	url	*Nullable.* A URL provided by the user in association with their profile.
	entities	Entities which have been parsed out of the url or description fields defined by the user.
	time_zone	*Nullable*. A string describing the Time Zone this user declares themselves within.
	utc_offset	*Nullable*. The offset from GMT/UTC in seconds.
	lang	The BCP 47 code for the user's self-declared user interface language. May or may not have anything to do with the content of their Tweets.
	is_translator	When true, indicates that the user is a participant in Twitter's translator community
	verified	When true, indicates that the user has a verified account. Verified accounts are usually those of public figures and celebrities.
	contributors_enabled	Indicates that the user has an account with "contributor mode" enabled, allowing for Tweets issued by the user to be co-authored by another account. Rarely true.
	geo-enabled	When true, indicates that the user has enabled the possibility of geotagging their Tweets. This field must be true for the current user to attach geographic data
	location	*Nullable*. The user-defined location for this account's profile. Not necessarily a location nor parseable.

It should be noted that most *Nullable* fields are usually empty, especially those that rely on the user's settings, such as "time_zone", "place", "coordinates", "location"… etc. Most users prefer keeping a degree of anonymity; they might not want people knowing who they are, where they're from or where they're posting from, what they do want is for people to read their tweets and know what they're thinking. This is usually the case with most users in the Arab region.

3.2 User Accounts Classification

Not all Twitter accounts are the same, they are highly diverse, but can be categorized into three obvious types of accounts: personal accounts, each belonging to genuine individuals; managed accounts, belonging to a group of people or a corporation; and finally, bot-controlled accounts, often referred to as twitterbots, which is an automated system administered by a computer program, which generates tweets.

When it comes to ***personal*** accounts, different users exhibit different behaviors. Naaman et al. (2010) categorized active users based on the type of messages that they typically post. The analysis resulted in two clusters, which were labeled as "Informers" (20% of users) and "Meformers". Meformers typically posts messages relating to themselves or their thoughts, whereas Informers post messages that are informative in nature. As for ***managed*** accounts, they arise because corporations, organizations, or even just a group of people with a common interest and cause would create a single Twitter account and appear as one. Sometimes even high ranking officials and public figures would have a dedicated team handle their account and to post tweets on their behalf. The tweets posted on these types of accounts often do not express the views or opinions of an individual, but of the group as a whole. ***Twitterbots*** also come in various forms. Aside from the fact that some may be fake or serve as spam, there are countless automated accounts that post news headlines, weather updates and even sports scores, while others may post at-reply messages in response to tweets that include a certain word or phrase, and some automatically retweet posts including a certain word or phrase.

We want to include only the personal accounts to be scored and ranked for the influential members' detection. Managed and Twitterbot accounts do not exhibit the influence we are searching for. Candidate accounts should be those of active genuine individual users.

Two account classification methods are proposed; a manual and an automated one. The manual approach simply consists of a manually assembled list of non-personal accounts to exclude if encountered. As for automated account classification, a machine learning approach seems appropriate and straightforward.

Statistical learning theory concerns the problem of choosing desired functions on the basis of empirical data. Support Vector Machines is the most prominent approach among modern results in this field (Kokash, 2005). SVMs support classification tasks based on the concept of optimal separator. The classification problem can be stated as a problem of data set separation into classes by the functions which are induced from available instances. The objective is to separate classes by the hyper-plane without errors and maximize the distance between the closest vectors to the separating hyper-plane.

Travis and Faisal (2013) studied the behavior of different types of Twitter accounts, examining the inter-tweet delay and the tweet time distribution for each class. The Twitter activity analysis showed that there are different patterns of tweeting activity across the Twitter account classes, suggesting that automated classification of account holders is possible without having to parse the content of the tweets.

We want to classify the accounts using a set of the basic user account features. For each unique user account in a collection we have the following relevant information, which we'll be using as attributes:

- **followers_count**: the number of followers the account currently has.
- **friends_count**: the number of users this account is following (AKA their "followings").
- **listed_count**: the number of public lists that this user is a member of.
- **favorites_count**: the number of tweets this user has favorited in the account's lifetime.
- The user's **activity rate**, calculated from the "**statuses_count**", which is the number of tweets (including retweets) issued by the user, and the account "**created_at**" date.

The goal of SVM is to produce a model, using some training data, which predicts the target values of the data given only the data attributes. LIBSVM (Chang and Lin, 2011) is currently one of the most widely used SVM software. A typical use of LIBSVM involves two steps: first, training a data set to obtain a model and second, using the model to predict information of a testing data set.

The following are the steps by which the SVM model was prepared:

1. The annotated data was converted into LIBSVM format which contains only numerical values. The general format of a record in the file is:

   [label] [index1]:[value1] [index2]:[value2] ...

   label : Sometimes referred to as 'class', the class (or set) of your classification, usually represented by integers. *index*: Ordered indexes, usually continuous integers. *value*: The data for training, usually lots of real (floating point) numbers.

   The features provided per user are: the number of followers, the number of followings, the number of lists the user is a member of, the number of favorite tweets and the user's activity rate.
2. Linear scaling was carried out on the data to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges and to avoid numerical difficulties during the calculation.
3. 10-fold cross validation was carried out.

After the SVM model is trained, test data is classified using both classification s; the manual and the SVM. We compare between them to find out which is the most effective and reliable in detecting the accounts to include in the ranking model. For each classification we set up a confusion matrix where each column in the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The confusion matrix allows more

detailed analysis than just accuracy, which is not a reliable metric for the performance of a classifier.

Table 2: sample confusion matrix

		Predicted	
		Personal	Non-Personal
Actual	Personal	True Positive instances	False Negative instances
	Non-Personal	False Positive instances	True Negative instances

Where the accuracy, precision, recall and specificity are calculated as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall\ = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

3.3 Detecting the topic-specific Influential Users

From a collection of topic-specific tweets retrieved, using a specific keyword, we want to detect the topic-specific influential users in the collection by developing a model that relies on the Twitter features that accompany the queried tweets. We first carry out feature selection; select the set of features we presume would add value to the model and help reach our goal. We then use these selected features in our experiments to develop a ranking model. In each of the experiments, the developed model will be evaluated, in order to decide on which is the best approach for reaching our goal.

3.3.1 Feature Selection

We want to use the tweet features to develop a model for ranking the users according to influence. In order to consciously use the features, we need to test their dependencies and see how

they relate to one another. Using a collection of queried tweets we get the correlation values between each of the relevant features related to the user tweets; dependent features will show high correlation values.

The following are the features extracted or generated, from the list of features in Table 1, which are relevant to our research and may be used as parameters in a ranking model formula. Based on our analysis, these are the features used by most researchers to detect influential users:

- **Statuses count**: The number of tweets (including retweets) issued by the user. It could be used as an indicator of the user's activity.
- **Account age (in days)**: The number of days since the account was created on Twitter.
- **The user's average daily activity rate:** the average of how many times a day the user posted. A better and clearer indication of the user's activity. Since Romero et al. (2010) found that influential individuals are often highly active users.

$$Average\ activity\ rate = \frac{statuses\ count}{account\ age}$$

- **Account age_activity combination**: Combining the Account's age and average activity rate, since according to Zhou et al. (2009), the earlier they join and more active they are, the more likely they will be considered as leaders.

$$Account\ age\ activity\ combo = 0.5 * Normalized\ account\ age + 0.5 * Normalized\ average\ activity\ rate$$

- **Followers count:** The number of followers an account currently has is an approximate indicator to the size of that user's audience. According to Keller and Berry (2003), one is influential if they are recognized by fellow citizens. Very few followers would mean that the user's message wouldn't reach many. The more the followers the more impact the user may have on Twitter.

- **TFF Ratio (Twitter Follower-Friend Ratio):** The ratio of a user's followers count to friends count. Bigonha and Cardoso (2010) used use this metric, combined with others, to identify influential users, considering users with higher TFF Ratio as more relevant. According to Donaldson and Hounshell (2009), a ratio of less than 1.0 indicates that the user is probably a listener and is seeking knowledge. A ratio of around 1.0 means that the user is respected among their peers, many think that a ratio of around 1.0 is the best – the user is listening and being listened to. A ratio of 2.0 or above shows that the user is a popular person and people want to hear what they have to say. Finally, a ratio of 10 or higher indicates that the user is either a Rock Star in their field or they are an elitist and cannot be bothered by Twitter's chatter.
- **Listed count:** The number of public lists that this user is a member of. A Twitter List is another method with which one can 'follow' Twitter users. Twitter Lists allows users to categorize other people on Twitter, assigning them to groups which will have their very own feed. Viewing a list timeline will show the user a stream of tweets from only the users on that list. These lists act as a filter on Twitter, making sure that the tweets of those who are of interest are picked up, away from the regular stream of tweets that sometimes renders the main Twitter Home Feed a meaningless, discursive babble. Quercia et al. (2011) claimed that those who are often listed in others' lists are usually the highly read users.
- **Collection tweets count**: The number of tweets in the queried collections posted by the user. It could be used as an indicator of a user's involvement and productivity in a particular topic (the queried topic). According to Akritidis et al. (2011), a user is considered to be productive if they had posted several posts recently and although productivity and influence do not coincide, there is quite a strong relationship between them, and therefore should somehow be taken into account.

- **Average Retweet count**: The average number of times a user's tweet(s) has been retweeted. To retweet is to republish something another user has written; rebroadcasting it to the followers, exposing them to what is believed to be valuable and/or entertaining content; it means that the message is being amplified. Retweets, which is one of the most popular features used in the literature, suggest that the tweet has resonated enough with someone that it encouraged them to pass it along and share it with their followers; the most obvious measure of a tweet's popularity. Acting as reinforcement to the message, it can be viewed as an endorsement of quality and a reflection of the user's ability to generate content with pass along value that got recognized by others.
- **Average Favorited count**: The average number of times a user's tweet(s) was marked as favorite by others. Twitter Favorites were first used solely to bookmark tweets a user wanted to read later. But recently, Favorites are used similar to the "like" button on Facebook. Even though scarcely used in comparison to retweets, users would favorite tweets to express, *I like what you're saying here*, or to answer a yes/no question in the affirmative. Nonetheless, it suggests that the tweet has resonated enough with someone that they want to bookmark it or let the author know that they liked i**t**
- **Average Tweet age**: The average age of the user's collection tweets, in minutes. The age of a tweet may be looked at from two different perspectives. The first is novelty, which was suggested by Keller and Berry (2003). Authors of the older collection tweets may be viewed as those who first started mentioning or discussing the topic. The second is taking into account the rapid changes; a topic being discussed and is of importance may be totally outdated in less than a couple of days, so for older posts to be kept alive, through retweets, is indication of its importance and ongoing effect on people.
- **Average Retweet frequency**: Factoring in time; the average tweet age, this feature is the average number of times a tweet was retweeted, per minute. It reflects the rate at which the

message spread across the network. For example, there may be a two day old tweet and a five minute old tweet, both with the same number of retweets, the retweet frequency is what would set them apart and show that the latter gained more recognition.

$$Average\ retweet\ frequency = \frac{avearge\ retweets\ count}{average\ tweet\ age + 1}$$

(The addition of 1 to the "average tweet age" in the denominator is to avoid division by zero. There are some cases where the tweets may have been posted in the same minute of their retrieval, resulting in their age being zero minutes.)

The use of the average of some of the features (Average Retweet Count, Average Favorited Count and Average Tweet age) is due to some of the users having more than one tweet within the collection. The average is used a representative of the feature for the user.

3.3.2 Ranking Users

The objective is to devise a method, using the selected features, which would rank the users according to topic-specific influence. Based on our understanding of the features, on a collection of topic-specific tweets, we first rank the users according to each of the selected features independently to see the effect of each of the selected features on the users ranking. We then develop and experiment with two different user ranking methods. In the first method, we use equations combining the best of the selected features. In the second, the users ranked in the top 50 according to each of the selected features were divided into sets according to their appearance frequency in the lists. By evaluating the different results we determine the method best results in the most satisfactory topic-specific user ranking. Finally we verify the effectiveness of the best of the ranking methods using a number of different collections.

3.3.3 The Model Evaluation Method

As mentioned in the previous chapter, there seems to be no training and testing data to evaluate the efficiency of a proposed approach. The absence of ground truth about influential bloggers presents another challenge. The key issue is how to find a reasonable reference point.

As briefly reviewed in the previous chapter, there are several online tools which analyze a person's Twitter presence. However, these measures are often not very definitive and not entirely reliable. The most popular one that was used in some other research papers is Klout. An API for Klout is available and easy to use to retrieve users' Klout scores. The use of these scores as a reference for evaluation was a tempting idea. From Twitter, Klout measures influence by using data points, such as following count, follower count, retweets, list memberships, how many spam or dead accounts are following a user, how influential the people who retweet are, and unique mentions. However we decided against using it as an evaluation reference point. Gaffney and Puschmann (2012) argue that the Klout score's lack of transparency undermines its status as a trustworthy metric. They also argue that Klout and similar services "gamify" the notion of influence in ways that encourage competitive behavior in ways which are detrimental to the quality of measurement in a scientific sense. Also, having carried out a few experiments with which we used Klout for evaluation in (Shalaby and Rafea, 2013), we later found that the user's Klout score to be highly affected by the number of followers in general, and was not indicative to the topic-specific influence we are interested in.

The lack of an obvious reference point with accurate information regarding influential users on Twitter got us to resort to a manual evaluation approach. Manual evaluation, however, is very labor intensive and can only be carried out on a limited scale. For a list of ranked influential users, manual evaluation of the users could give indication of how good the ranking model is. However, despite it being a challenging task, we find it to be the most reliable and suitable to our search for the topic-specific influential users.

For the ranking methods' development carried out on a specific queried collection of tweets, we assemble a list of the topic-specific influential users in that collection. The list is assembled in order to be used as a reliable reference for the methods' evaluation. The list of influential users is assembled by ranking the users according to each of the independent features, and studying each of the users ranked in the top 50 by each of the features. By studying each of the users, we determine the most influential users in the collection. As for the verification experiments carried out on a number of different collections of queried tweets, we only study the candidate users proposed by the methods as influential.

To manually determine whether a user is influential or not, we first read the user's tweets; the collection tweets that put the user in that ranking by the method, and judge the content of the tweet based on its relevance to the queried topic, how well written it is and the message it conveys.

Then by going to the user's Twitter profile page we first check out the user's mini-biography. The Twitter bio is the first thing people will read when they view someone's Twitter page. It is one of the decisive factors when deciding whether to follow or not follow that person. A good bio would often include a few critical keywords that would describe the user and the nature of the posts. From the bio we find out how the user portrays themselves, their interests, and in some cases, who the user is; their name and occupation; if they're a public figure or celebrity, writer, journalist, activist… etc. The bio may also give indication how the account is being used; for the user's personal expression and life logging, or in support of a certain cause. Also, some users would include a web-link, for example, to their Facebook page, an official webpage, a personal blog or Youtube channel. Statistics reported by Zarrella (2009) show that Twitter profiles that contain a bio will attract eight times more followers on average than users without a bio and users with a web-link have over 7.5 times as many as users without. The web-

links often provided in a user's bio may be useful in providing further information to assist the evaluation.

In the user's profile page we go over a few of the user's recent activity; observing the ratio of original posts to retweets, favoring users with more original posts than retweets. We also observe their content, writing style, topic interests; their consistency and relevance to the topic-specific queried tweets, and also the target audience; whether their posts address their friends and acquaintances or the general masses of readers. Also, we view some of the correspondences with other users and how users reply to their posts, to see how they interact with their readers and how their readers react and respond; whether positively and supportive or in disagreement.

The user's posts and conversations should not be too self-centered and self-involved, but commonly discuss trending topics and issues of common interest. Even regular individuals may be influential, but they should be actively and personally involved with a wider audience than just friends and acquaintances, posting original content that conveys a purpose or useful message.

From 1221 unique users in the queried collection used for the ranking methods' development, we identified and listed 31 influential users. We use this list to evaluate the different ranking methods carried out on that collection. We measure how many of the users from the annotated list made it to the top ranking according to the model and calculate the precision, which represents the fraction of the users that are considered influential.

Besides the use of precision as an evaluation metric for each of the ranking methods, we use significance testing (T-test). A T-test's statistical significance indicates whether or not the difference between two groups' averages most likely reflects a "real" difference. We use the T-test to see if the difference or improvement of the values between two ranking methods is statistically significant or not.

The test of significance begins with a *null hypothesis* which represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. Then there is the *alternative hypothesis* which is what the statistical hypothesis test is set up to establish. Once the test has been carried out, the final conclusion is given in terms of the null hypothesis; it is either rejected or not. If the null hypothesis is not rejected, this does not necessarily mean that the null hypothesis is true; it only suggests that there is not sufficient evidence against the null hypothesis. Rejecting the null hypothesis only suggests that the alternative hypothesis may be true.

We carry out the T-test using the two-sample t-test (ttest2) in the MATLAB Statistics Toolbox. The function $h = ttest2(x, y)$ returns a test decision for the null hypothesis that the data in vectors x and y comes from independent samples from normal distributions with equal means and equal but unknown variances. The alternative hypothesis is that the data in vectors x and y comes from populations with unequal means. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

3.4 Using Statistical Language Modeling (SLM) for Linguistic analysis of the Tweet text

Eloquence was one of the properties of an influential users set out by Keller and Berry (2003). Kiciman (2010) demonstrated the importance of linguistic style variations in Twitter by examining differences in language models in Twitter posts related to different metadata, and according to Quercia et al. (2011), different types of individuals use language differently in their posts and they have found that it is linked to social influence. In order to find out if the tweet language may be used as an indicator of influence. With the assumption that highly retweeted users are more likely to be influential, we test the use of Statistical Language Modeling to measure the quality of the tweet text and if it may be related to the retweets count of a tweet.

A Statistical Language Model is simply a probability distribution $P(s)$ over all possible sentences s (Rosenfeld, 2000). Statistical language modeling (SLM) is the science of building models that estimate the prior probabilities of word strings. It is crucial for applications in natural language technology and other areas where sequences of discrete objects play a role. These include speech recognition, machine translation, document classification and routing, optical character recognition, information retrieval, handwriting recognition, spelling correction, and many more. SLM employs statistical estimation s using language training data. The most successful SLM s use very little knowledge of what language really is. The most popular language models, the N-gram models, takes no advantage of the fact that what is being modeled is language, it may as well be a sequence of arbitrary symbols, with no deep structure, intention or thought behind them (Rosenfeld, 2000). Its basic idea is to consider the structure of a corpus as the probability of different words occurring alone or occurring in a sequence.

The N-gram models estimate the probability of each word given prior context. An N-gram model uses only $N-1$ words of prior context. So in an N-gram model, the underlying assumption is:

$$P\left(w_i \middle| w_1^{i-1}\right) = P(w_i | w_{i-(n-1)}, .., w_{i-1})$$

where the conditional probability is calculated from the N-gram frequency counts of word sequences from the training corpus:

$$P\left(w_i \middle| w_{i-(n-1)}, .., w_{i-1}\right) = \frac{c(w_{i-(n-1)}, .., w_{i-1}, w_i)}{c(w_{i-(n-1)}, .., w_{i-1})}$$

where $c(w)$ stands for the count of occurrences of the substring w.

However, the n-gram probabilities are not derived directly from the frequency counts. A smoothing process is used for making the model more robust to phenomena that were not

observed in the training data by assigning some of the total probability mass to unseen words or N-grams.

We make use of the SRI Language Modeling Toolkit (SRILM), which is an open source software toolkit for building and evaluating statistical language modeling and related tasks. Most of the SLM types it supports are based on N-gram statistics, including the standard back-off models, with an array of standard smoothing algorithms (Stolcke et al., 2011).

3.4.1 Building the training corpus

For the training corpus, a number of queries are carried out on a variety of highly discussed topics. We want to create a training corpus of the popular tweets; the tweets that resonated with enough users that they got acknowledged in the form of many retweets. So any tweet with a retweet count less than 20 will not be included in the training corpus of our language model.

Tweet text is known to often contain information besides the actual text message being posted. For example, the message could be a retweet and would thus contain "RT @username:" or, the tweet could be a reply or just a message to another user, in both these cases @username will be included so that other user is sure to see it. Using URLs to reference material on the internet is also quite popular. Symbols and different characters are also occasionally used in the text, especially "#", the hash used for tagging the tweet with keywords (topics) it may be relevant to.

Regardless of all that may be included in the tweet, the most important is the actual message the user originally intended to broadcast. That is the text we are interested in extracting and evaluating. That is the text we'll be building our SLM training corpus with. The text preprocessing is actually quite simple in our case. Remove user mentions (@username), remove

URLs, and remove non alphanumeric symbols and non-Arabic characters. This would leave us with plain Arabic text.

3.4.2 Building the Statistical Language Model using SRILM

The main purpose of SRILM is to support language model **estimation** and **evaluation**. Estimation means the creation of a model from training data; evaluation means computing the probability of a test corpus, conventionally expressed as the test set perplexity. SRILM by itself performs no text conditioning and treats everything between white spaces as a word. The functions to accomplish these two purposes are named ***ngram-count*** and ***ngram***, respectively (Stolcke, 2002).

3.4.2.1 Model Estimation

To create a model from the training corpus, three main steps are carried out (Chen, 2008):

1. Generate the n-gram count file from the corpus
2. Train the language model from the n-gram count file
3. Calculate the test data perplexity using the trained language model

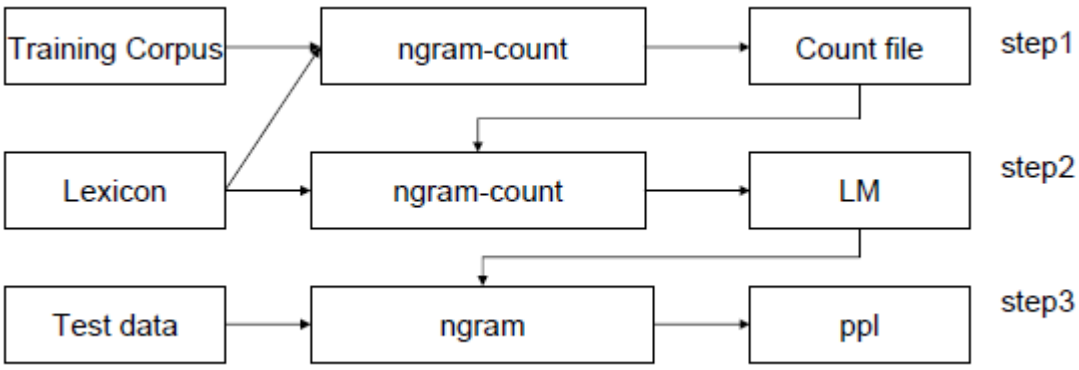

Figure 1: SRILM workflow

Any lexicon file may be used with the ***ngram-count*** function. Since we didn't have one, we generated our lexicon from the training corpus. That ***ngram-count*** function has an option which generates a lexicon file from an input text file.

Using the default toolkit options, we use a standard SLM; trigram with Good-Turing discounting and Katz backoff for smoothing, to capture the style of what the popular tweets tend to look like, so as to help us detect popular tweets based on the language used.

3.4.2.2 Model Evaluation

As stated by Rosenfeld (2000), to assess the quality of a given language modeling , the likelihood of new data is most commonly used. The average log likelihood of a new random sample is given by:

$$Average\ Log\ Likelihood\ (D|M) = \frac{1}{n}\sum_{i} \log P_M(D_i)$$

where $D = \{D_1, D_2, \dots, D_n\}$ is the new data sample, and M is the given language model. This latter quantity can be viewed as an empirical estimate of the cross entropy of the true, but unknown data distribution P with regard to the model distribution P_M:

$$cross\ entropy(P; P_M) = -\sum_{D} P(D).\log P_M(D)$$

The performance of the language model is often reported in terms of perplexity (Bahl et al., 1977):

$$perplexity(P;\ P_M) = 2^{cross\ entropy(P;P_M)}$$

Perplexity is the preferred metric for practical language model construction. Perplexity can be interpreted as the (geometric) average branching factor of the language according to the model. It is a function of both the language and the model. When considered a function of the

model, it measures how good the model is (the better the model, the lower the perplexity). When considered a function of the language, it estimates the entropy, or complexity, of that language (Rosenfeld, 2000). The lower the perplexity value, the closer the language is to the model.

Our hypothesis is that if a negative correlation does exist between the tweet text perplexity and the retweet count, then maybe the perplexity measure can be used to predict the popular tweets that have a high probability of being recognized and retweeted. We are going to carry out a few experiments; we first want to see the how the users' tweets' average perplexity values relate with the other user features, we do that by calculating the correlation values. The effect of increasing the training data size is also investigated. We then investigate the effect of a user's tweets' perplexity on their influence and if there is a relation between the two. After further examination of the influential users' perplexity ranges and tweets' work count, we try to incorporate the perplexity into one of the ranking s to see the effect it might have on the outcome.

CHAPTER 4

Detecting Influential Users using the Twitter Features

In this chapter we first investigate the effectiveness of using SVM for user account classification. Secondly, for influential users' detection, we carry out feature selection before deciding on which features to use to develop the users' ranking model. From the selected features set we use the independent features to experiment with different ranking models to settle on the that results in the most satisfactory topic-specific influential user ranking, so we may rely on it to detect the most influential users in a collection.

4.1 User Accounts Classification

The objective of this experiment is to investigate the effectiveness of SVM for user account classification using some basic numeric account features, and to find out which of the two approaches is more reliable; the manually assembled list of accounts to filter out or the SVM. We are interested in classifying the personal user accounts from which we want to detect the influential users.

4.1.1 Data Description

To make up the SVM training dataset, tweets form 6 different queries were used. From that collection, 5471 unique users were identified and manually annotated as personal or non-personal.

As for the SVM testing datasets, two sets were prepared; one for each of the tests carried out. For the first, a test dataset of 1221 annotated user accounts was prepared. The users were extracted from tweets from a single query; one relevant to the current events in Egypt. As for the second, another test dataset of 1092 annotated user accounts was prepared. The users were also extracted from tweets for a single query. However, for this second test, the query domain was

different. The tweets are about Cairo in general, and this time the query was in English instead of Arabic as we usually do.

4.1.2 Method

The Support Vector Machine, prepared by 10-fold cross validation, produced an accuracy of **88.26%**, and the following confusion matrix shown in Table 3.

Table 3: the 10-fold cross validation confusion matrix

		Predicted		Accuracy = 0.882
		Personal	Non-Personal	Precision = 0.989
Actual	Personal	4687	590	Recall = 0.888
	Non-Personal	53	141	Specificity =0.727

Before developing the SVM, we used to use a list of manually assembled collection of non-personal accounts as a reference to differentiate the personal from the non-personal accounts; if the account was found on the list, then it was to be filtered out.

We want to compare between the effectiveness of account differentiation between using SVM and a manually assembled a list of accounts. To compare with the SVM classification, we have two lists. The first list consists of 105 non-personal accounts and was last updated in June 2012, during which we were carrying out some preliminary experiments. The second list consists of 681 non-personal accounts and was last updated in October 2013. It consists of the non-personal accounts we encountered until that time. In both cases the accounts in the lists were among those discussing a range of events and issues related to the political scene in Egypt in Arabic written tweets.

We first compare between the classification accuracy of the SVM and the lists. Then knowing that the lists are domain specific, since they were assembled while analyzing tweets

about local, politically related, events in Egypt sometime during 2012 and 2013, we investigate how a change in domain may affect the accuracies of both the list and the SVM outcomes.

4.1.3 Results

We first compare between the classification accuracy of the SVM and the lists. Using the test dataset of 1221 annotated user accounts prepared for this experiment, each of the Tables 4, 5 and 6, show the confusion matrices of the data having been classified by the SVM, the October 2013 list and the June 2012 list, respectively.

Table 4: Confusion matrix for the SVM classification

		Predicted		
		Personal	Non-Personal	Accuracy = 0.901 Precision = 0.905 Recall = 0.991 Specificity =0.265
Actual	Personal	1060	10	
	Non-Personal	111	40	

Table 5: Confusion matrix for classification using the October 2013 list

		Predicted		
		Personal	Non-Personal	Accuracy = 0.932 Precision = 0.928 Recall = 1 Specificity =0.45
Actual	Personal	1070	0	
	Non-Personal	83	68	

Table 6: Confusion matrix for classification using the June 2012 list

		Predicted		
		Personal	Non-Personal	Accuracy = 0.881 Precision = 0.88 Recall = 1 Specificity =0.04
Actual	Personal	1070	0	
	Non-Personal	145	6	

Then using the second test dataset of 1092 annotated user accounts prepared for this test, having changed the query domain, each of the Tables 7, 8 and 9, show the confusion matrices of the data having been classified by the SVM, the October 2013 list and the June 2012 list, respectively.

Table 7: Confusion matrix for the SVM classification

		Predicted		Accuracy = 0.943 Precision = 0.946 Recall = 0.994 Specificity =0.462
		Personal	Non-Personal	
Actual	Personal	982	6	
	Non-Personal	56	48	

Table 8: Confusion matrix for classification using the October 2013 list

		Predicted		Accuracy = 0.909 Precision = 0.909 Recall = 1 Specificity =0.048
		Personal	Non-Personal	
Actual	Personal	988	0	
	Non-Personal	99	5	

Table 9: Confusion matrix for classification using the June 2012 list

		Predicted		Accuracy = 0.906 Precision =0.906 Recall = 1 Specificity =0.01
		Personal	Non-Personal	
Actual	Personal	988	0	
	Non-Personal	103	1	

4.1.4 Discussion

Despite the significant data diversity in each of the different types of accounts, the 10-fold cross validation was able to classify the personal accounts with 0.882 accuracy, 0.989 precision and 0.888 recall.

Recall, which is the fraction of relevant instances that got predicted, should have been the measure for evaluating the system's performance on the test data. However, as may be seen from the results in Tables 5, 6, 8 and 9, the recall value is 1. This is due to the use of the lists; there are no false negatives at all, and false positives are not accounted for in the definition of recall. Therefore, it alone cannot be used to determine whether a test is useful in practice. This led us to resort to both precision and specificity as performance measures to properly evaluate our test results.

Even though the manually assembled October 2013 list may have produced the better accuracy, precision and specificity, the results of the older June 2012 list show that over time the

list performance deteriorates. In order to maintain good performance, the list would require continuous updating. Also, the lists are domain specific since they were assembled during our investigation of tweets from a certain domain. When we changed the domain of test dataset, the performance of the lists was surpassed by the SVM. The October 2013 list performance decreased and the older list was rendered useless with a very low specificity value of 0.01; capturing almost none of the non-personal accounts.

Even though the performance measures of the SVM were close to those of the lists, the SVM is domain independent. The SVM relies on numbers which reflect some user account behavior in general, so even if there was some misclassification, it is reliable and consistent despite the error margin.

4.2 Feature Selection

The objective of this section is to decide on the features to use in the ranking model. We study the correlation values between the relevant tweets features associated with each user. If two features are highly correlated, it is redundant to use both in the user ranking model; we use one or the other.

4.2.1 Data Description

From six of the retrieved queries, the same queries used to train the SVM model. We extract the necessary information from the 10,539 tweets and assemble a list of 5471 users; the tweet authors. Each user is associated with the following selected features:

- Statuses count
- Account age (in days)
- The user's average daily activity rate
- Account age_activity combination

- Followers count
- TFF Ratio (Twitter Follower-Friend Ratio)
- Listed count
- Collection tweets count
- Average Retweet count
- Average Favorited count
- Average Tweet age (in minutes)
- Average Retweet frequency (per minute)

4.2.2 Method

The correlation values between each of the selected features are calculated to shed light on dependencies between the different user features. We used the Microsoft Excel built-in correlation function to carry out the calculations. According to (Microsoft, 2014), the function $correl(X, Y)$ returns the correlation coefficient of the array X and array Y cell ranges. For a set of observations (x_1, y_1), (x_2, y_2),… (x_n, y_n), the equation for computing the correlation coefficient is given by:

$$correl(X, Y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the array X and array Y cell ranges.

4.2.3 Results

Table 10 shows the correlation values between each of the relevant user features.

Table 10: The correlation values between each of the features

	Statuses Count	Account Age	Activity Rate	Age +Activity combo	Followers	TFF Ratio	Listed	Collection Tweets Count	Avg Retweet Count	Avg Favorite Count	Avg Tweet Age
Account Age	0.251	.									
Activity Rate	**0.653**	-0.120	.								
Age +Activity combo	**0.516**	**0.909**	0.300	.							
Followers	0.145	0.125	0.042	0.138	.						
TFF Ratio	0.158	0.057	0.068	0.082	0.410	.					
Listed	0.196	0.163	0.057	0.179	**0.957**	0.480	.				
Collection Tweets Count	0.356	0.001	0.388	0.163	0.078	0.080	0.101	.			
Avg Retweet Count	0.006	0.062	-0.017	0.052	0.478	0.125	0.454	-0.005	.		
Avg Favorite Count	0.000	0.068	-0.023	0.056	0.468	0.168	0.464	-0.003	**0.841**	.	
Avg Tweet Age	0.004	0.030	-0.014	0.023	0.081	0.132	0.087	-0.010	0.134	0.144	.
Retweet Frequency	0.026	0.036	0.008	0.038	0.250	0.027	0.260	-0.017	0.178	0.143	-0.014

4.2.4 Discussion

From Table 10 we can see that there are a few high correlations between some features. The highest correlation value, 0.957, is between users' Followers count and Listed count. This may be due to popular users, with high Followers counts, also being highly read users with high Listed counts, and vice versa. There is also a high correlation of 0.841 between the Average Retweets count and the Average Favorites count. The dependency between the Retweets and Favorites counts may be due to good tweets getting recognition from other users in the form of both retweets and favorites, and vice versa; little or no recognition to the not so impressive tweets. An expected high correlation exists between the users' Statuses counts and their Average Activity rates, probably since one is derived from the other.

As previously mentioned, if two features are highly correlated, it is redundant to use both in the user ranking model; we use one or the other. With a correlation threshold of 0.5, we decided to focus on Retweets count over the Favorited count. Retweets are more popular and frequently used than Favorites and they have a bigger impact; by spreading the message to a

bigger audience, unlike the favorite option, which is just between the two users. We also decided to focus on the Followers count over the Listed count, since the Listed count relies greatly on the other users' personal preferences; whether they organize the users they follow in lists or not and for what reasons, in addition to the fact that some users do not use lists at all. Between a user's Average Activity rate and Statuses count, we decided to focus on the Average Activity rate as a more accurate representation of a user's activity on Twitter.

Finally, the following is the list of features that we'll be experimenting with to detect the influential users:

- *Feature 1:* The user's average daily activity rate
- *Feature 2:* Account age_activity combination
- *Feature 3:* Followers count
- *Feature 4:* TFF Ratio (Twitter Follower-Friend Ratio)
- *Feature 5:* Collection tweets count
- *Feature 6:* Average Retweet count
- *Feature 7:* Average Tweet age (in minutes)
- *Feature 8:* Average Retweet frequency (per minute)

4.3 Ranking Users: according to each of the selected features independently

The objective of this experiment is to see the effect of ranking the users according to each of the selected features independently, and deciding which of the features rank the influential users best.

4.3.1 Data Description

For the ranking model experiments we decided to use just one of the tweets collections, which was a query on Novermber 5th, 2013, with the words "باسم يوسف", from which we extracted

a collection of 1221 unique users who posted tweets commenting on the cancellation of a popular TV show the prior weekend. The comments spanned around a 3 hour time window.

Account classification was carried out on the collection users. The personal accounts to include in our experiments were identified by the SVM with a precision of 0.905, so despite the classification efforts there was an error margin which could result in encountering some non-personal accounts in the ranking experiments. So for the following experiments the non-personal accounts were manually filtered, in attempt to avoid misguided outcomes or conclusions that may have been caused by misclassified accounts.

4.3.2 Method

For each of the independent features we settled on in section 4.2.4, we rank the users accordingly, in descending order, and evaluate their effectiveness in ranking the influential users.

For evaluating, we refer to the manually assembled list of 31 influential users, listed in Appendix A, according to which we calculate the rankings' precision values. To personally determine whether a user is influential or not, we first look at the user tweets in the queried collection. Then by going to the user's Twitter profile page we check out the user's mini-biography and view his/her recent activity; posts and retweets, in addition to observing the content, style and even some of his/her correspondences and replies to the posted tweets. What all the influential users have in common is that they are active on Twitter, frequently posting original content expressing their personal views and opinions that get recognition in the form of retweets or replies. Also, most of the accounts have a significant number of followers.

Finally, the T-test is carried out on the precision of the results of identifying influential users using pairs of features to see if the difference or precision improvement is statistically significant.

4.3.3 Results

The users of the collection are ranked according to each of the eight features. The top 50 users ranked according to each of the features may be seen in Tables 1, 2, 3, 4, 5, 6, 7 and 8 in Appendix B. For each feature, the precision at 10, 20, 30, 40 and 50 are calculated as can be seen in Table 11. Each row reflects one of the eight features according to which we ranked the users. For each of the top users groups, there is a *Count* column which contains the number of influential users found in that set, and a *Precision* column of the calculated precision value.

From Figure 2, which visual representation of the precision values in Table 11, we can see that Features 2, 3, 6 and 8 seem to have higher precision values that those of Features 1, 4, 5 and 6. In order to compare between the rankings of this experiment, the t-test was conducted between selected pairs, in Table 11, to test if the difference is statistically significant.

Table 11: Summary of the Precision values of experiment 4.3

Feature	Top 10		Top 20		Top 30		Top 40		Top 50	
	Count	Precision	Count	Precision	Count	Precision	Count	Precision	Count	Precision
1	1	**0.1**	1	**0.05**	2	**0.07**	2	**0.05**	3	**0.06**
2	4	**0.4**	9	**0.45**	9	**0.3**	9	**0.225**	9	**0.18**
3	7	**0.7**	12	**0.6**	16	**0.53**	18	**0.45**	21	**0.42**
4	3	**0.3**	8	**0.4**	11	**0.37**	15	**0.375**	17	**0.34**
5	2	**0.2**	4	**0.2**	7	**0.23**	7	**0.175**	9	**0.18**
6	6	**0.6**	9	**0.45**	13	**0.43**	14	**0.35**	18	**0.36**
7	2	**0.2**	4	**0.2**	4	**0.13**	5	**0.125**	7	**0.14**
8	4	**0.4**	9	**0.45**	14	**0.47**	16	**0.4**	17	**0.34**

Figure 2 provides visual representation of the precision values in Table 11; where for each of the features the precision at 10, 20, 30, 40 and 50 is plotted.

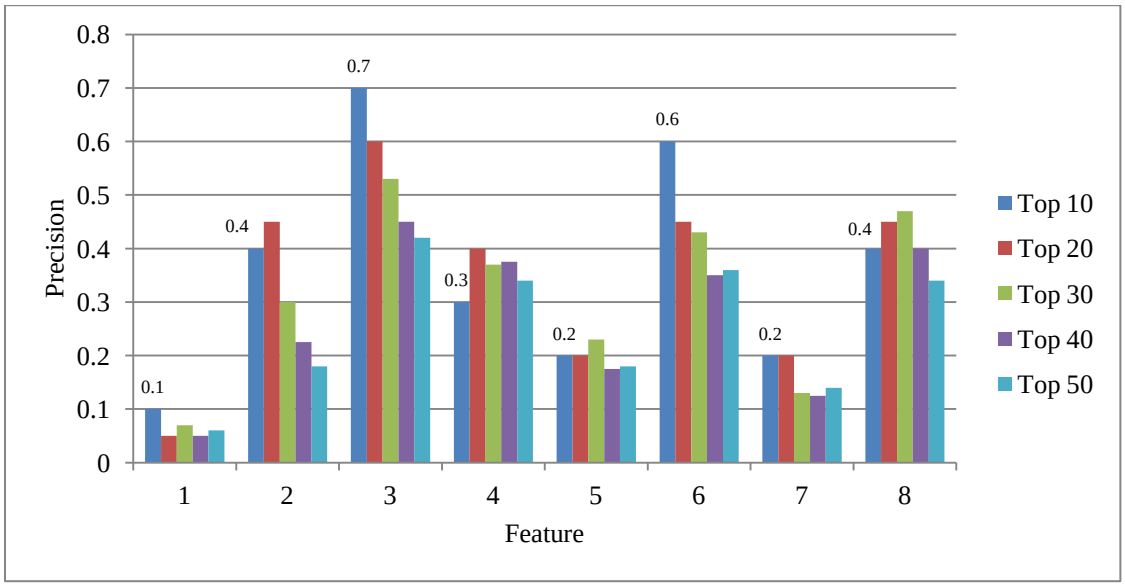

Figure 2: a visual of the precision values of experiment 4.3

Table 12 shows the two-sample t-test null hypothesis results; **rejected** or **not rejected**, for each of the pairs in the row and the column. The features in the rows are the better features, with higher precision values, and the features in the columns are did not result in good precision values for detecting influential users. The mean m and standard deviation s of the precision values of each of the features are stated in the table.

Table 12: the two-sample t-test null hypothesis results

	Precision values of Feature 1 **(m=0.06, s=0.02)**	Precision values of Feature 4 **(m=0.36, s=0.04)**	Precision values of Feature 5 **(m=0.2, s=0.02)**	Precision values of Feature 7 **(m=0.16, s=0.04)**
Precision values of Feature 2 **(m=0.2, s=0.1)**	rejected	Not rejected	Not rejected	rejected
Precision values of Feature 3 **(m=0.54, s=0.11)**	rejected	rejected	rejected	rejected
Precision values of Feature 6 **(m=0.44, s=0.1)**	rejected	Not rejected	rejected	rejected
Precision values of Feature 8 **(m=0.41, s=0.05)**	rejected	Not rejected	rejected	rejected

4.3.4 Discussion

Based on the precision at 10 values, which may be seen in Table 11, the feature with the highest precision was Feature 3 (Followers Count) at 0.7, followed by Feature 6 (Average Retweets Count) at 0.6. Feature 2 (Account Age_Activity Combination) and Feature 8 (Average Retweet Frequency) followed, both with 0.4 precision. After that, Feature 4 (TFF Ratio), with a 0.3 precision, and then Feature 5 (Collection Tweets Count) and Feature 7 (Average Tweets age), both with 0.2 precisions. Finally, with the lowest precision value of 0.1 is Feature 1 (Average Activity Rate).

The highest precision values obtained were of those ranked according to Feature 3 (Followers Count). The number of followers primarily reflects a user's popularity, and not necessarily their influence, since some users, despite their high followers counts, were not considered influential. For example, the public figures, such as writers, journalists, TV presenters and reporters, celebrities, government officials and politicians, are known by many people of the public and are recognized by them. People are often interested to read what they have to say and would like to be kept up to date with their activities and posts. The large audience size does not necessarily imply that they're all interested in everything being tweeted or that all tweets get the same amount of attention.

So despite the concept that popular users with large numbers of followers are not necessarily influential, it seems that influential users often have a considerable number of followers. So the number of followers should be taken into consideration as a contributing factor to a person's influence strength; the more the followers, the more the message is likely to spread, but it is not very reliable and cannot be used on its own. Also, some highly followed users are simply not influential, or may have been considered influential at some point in time, but cease to be anymore. For example, "Almoslemani", ranked fourth by the Followers Count in Table 3 in Appendix B; a reported, turned TV presenter, turned advisor for the president for media affairs,

has a very high followers count, however, his once possibly influential Twitter account has now become a series of news headlines and links to articles that either mention or quote him, and he no longer posts his personal opinions. So despite the large number of followers, this user is not considered influential in the Twittersphere.

The second highest precision values obtained were of those ranked according to Feature 6 (Average Retweets Count). The results show that not because a user managed to post a popular tweet or a few that got highly retweeted then that user may be considered influential. Despite their high retweets counts, some users were still not considered as influential. Nonetheless the retweet counts are a solid reliable measure to the amount of attention and response a tweet gets.

The lowest precision values were a result of ranking according to Feature 1 (Average Activity Rate). From the ranking, we found that the influential users are not the most active. Over 100 tweets and/or retweets a day is considered quite a lot, and may be regarded as spamming. It seems that the influential users are more selective and conscious of what they post on Twitter, which we found to be also reflected in Feature 5 (Collection Tweets Count); most of the influential users had posted one or two tweets in the collection, unlike some others who had posted up to 19 and 20 tweets. We found that both Features 1 and 5 cannot be used to detect the influential users in a collection, because despite the presence of a few influential users who do write a lot, most of the very highly active users are not influential. Also with a very low precision value is Feature 7 (Average Tweets age). We found out that the age of the tweet cannot be used as an indication of a user's influence. Another feature we found we cannot rely on to detect influential users is Feature 4 (TFF Ratio) with a low precision value of 0.3.

Combining the account's age with its average activity rate, creating Feature 2 (Account Age_Activity Combination), got some of the older accounts into the higher ranks, slightly

improving the ranking precision from 0.1 to 0.4. Even though the precision value is low it is still relatively better. Also with a precision of 0.4 is Feature 8 (Average Retweet Frequency).

As previously stated and may be seen in Figure 2 we can see that Features 2, 3, 6 and 8 seem to have higher precision values that those of Features 1, 4, 5 and 7. The t-test was conducted between selected the best and worst feature pairs to test if the difference is statistically significant. From the T-test null hypothesis results in Table 12, we can see that the precision values of Feature 3 are better than those of all four features; Features 1, 4, 5 and 7, with statistical significance. While for features 6 and 8, their values were found to be better than three of the four worst features with statistical significance. In case of Feature 2, it was found to be better than two of the four worst features with statistical significance. These results guided us into using the best four features in our next experiment where we investigate the effect of combining features on the users' ranking.

4.4 Ranking Users: combining the best features

The purpose of this experiment is to see the effect of combining the best features on the users' ranking.

4.4.1 Data Description

This experiment uses the same tweets collection used in the experiment in section 4.3, which was a query on Novermber 5th, 2013, with the words "باسم يوسف", from which we extracted a collection of 1221 unique users.

4.4.2 Method

According to experiment 4.3, the four features which resulted in the highest precision at 10 values, with statistical significance, for influential users are:

- *Feature 2:* Account Age_Activity Combination (**AAcombo**)

- *Feature 3:* Followers Count (**F**)
- *Feature 6:* Average Retweets Count (**RT**)
- *Feature 8:* Average Retweets Frequency (**RTfreq**)

We rank the users according to each of the following combined features scores; combining the best four, three and two features:

$$Score\ 1 = \frac{1}{4}(Normalized\ \boldsymbol{AAcombo} + Normalized\ \boldsymbol{F} + Normalized\ \boldsymbol{RT} + Normalized\ \boldsymbol{RTfreq})$$

$$Score\ 2 = \frac{1}{3}(Normalized\ \boldsymbol{AAcombo} + Normalized\ \boldsymbol{F} + Normalized\ \boldsymbol{RT})$$

$$Score\ 3 = \frac{1}{3}(Normalized\ \boldsymbol{F} + Normalized\ \boldsymbol{RT} + Normalized\ \boldsymbol{RTfreq})$$

$$Score\ 4 = \frac{1}{2}(Normalized\ \boldsymbol{F} + Normalized\ \boldsymbol{RT})$$

The rankings resulting from each of the above scores will also be evaluated according to the manually assembled list of influential users which may be found in Appendix A, and the T-test is conducted on some of the scores' precision values to see if the difference or precision improvement is statistically significant. Also, another T-test is conducted between the score with the highest mean precision value and the feature from experiment 4.3 also with the highest mean precision value.

4.4.3 Results

The users of the collection are ranked according to each of the four scores, the top 50 of which may be seen in Tables 9, 10, 11 and 12 in Appendix B. For each score, the precision at 10, 20, 30, 40 and 50 are calculated as can be seen in Table 13. Each row in the table reflects one of the scores according to which we ranked the users. For each of the top users groups, there is an

Influential Users Count row which contains the number of influential users found in that set, and a *Precision* row of the calculated precision value.

Table 13: the Precision calculated for each of the top groups ranked according to the scores

		Top 10	Top 20	Top 30	Top 40	Top 50
Score 1	Influential Users Count	6	11	15	17	19
	Precision	**0.6**	**0.55**	**0.5**	**0.425**	**0.38**
Score 2	Influential Users Count	7	11	13	16	17
	Precision	**0.7**	**0.55**	**0.43**	**0.4**	**0.34**
Score 3	Influential Users Count	8	12	17	19	20
	Precision	**0.8**	**0.6**	**0.57**	**0.475**	**0.4**
Score 4	Influential Users Count	9	12	15	19	20
	Precision	**0.9**	**0.6**	**0.5**	**0.475**	**0.4**

Figure 3 provides visual representation of the precision values in Table 13; where for each of the top 10, 20, 30, 40 and 50 groups the precision values for each of the score users rankings are plotted.

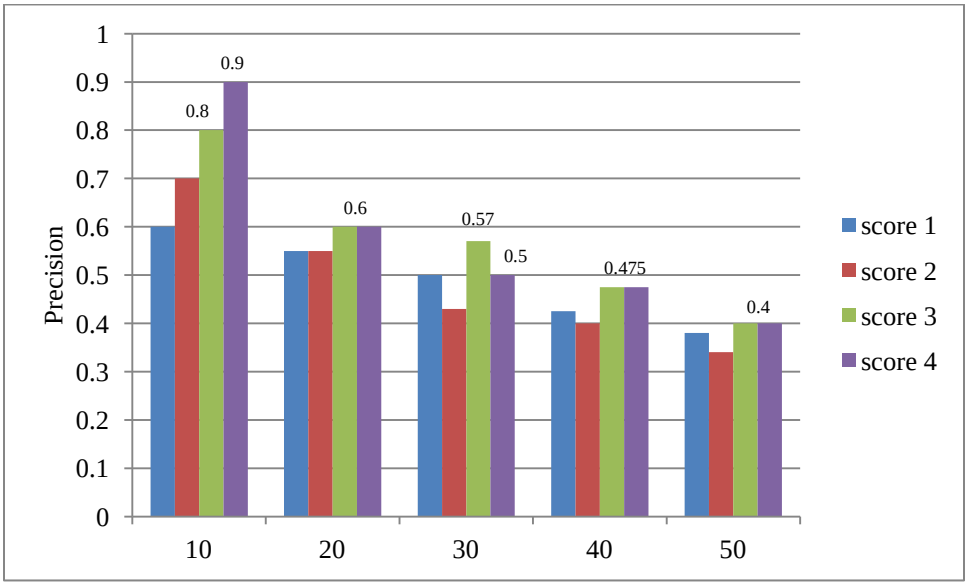

Figure 3: a visual of the precision values of experiment 4.4

From Figure 3 we can see that Score 3 and Score 4 seem to have higher precision values that those of Score 1 and Score 2. In order to compare between the rankings of this experiment, the T-test was conducted between selected pairs to test if the difference is statistically significant.

Table 14 shows the two-sample T-test null hypothesis results; **rejected** or **not rejected**, for each of the pairs in each row and each column. The mean m and standard deviation s of the precision values of each of the features are stated in the table.

Table 14: the two-sample t-test null hypothesis results

	Precision values of score 1 **(m=0.49, s=0.09)**	Precision values of score 2 **(m=0.48, s=0.14)**
Precision values of score 3 **(m=0.57, s=0.15)**	Not rejected	Not rejected
Precision values of score 4 **(m=0.58, s=0.2)**	Not rejected	Not rejected

The highest mean precision value is 0.58 of score 4, and from experiment 4.3, the highest mean precision value is 0.54 of Feature 3. The T-test was conducted in order to compare between their precision values, and the t-test null hypothesis was **not rejected**.

4.4.4 Discussion

The best features were combined into equations according to which the users were ranked according to the score. The top 10 rankings produced the best precision values; some of the best we've seen so far. The score of the equation combining the average retweets count, followers count and the average retweets frequency produced a precision at 10 of 0.8, and when just the average retweets count and the followers count are combined, it produced a precision at 10 of 0.9. Even though the results of the previous experiment, in section 4.3, show that those highly retweeted users are not necessarily with the highest numbers of followers, and those with the high number of followers do not all necessarily have the best or the most effective tweets in the collection, combining the features improved the results. This consolidates the hypothesis that influential users are recognized by many and that their posts resonate with other users and spread rapidly throughout the network.

The T-test was conducted to see if the difference between scores 3 and 4 over scores 1 and 2 is statistically significant. From the T-test null hypothesis results in Table 14, we can see that they did not reach statistical significance. Also, the difference between Score 4 precisions and the precisions of Feature 3 of the previous experiment did not reach statistical significance either.

4.5 Ranking Users: according to their appearance frequency when ranked by the features

The objective of this experiment is to see if the traits reflected by the eight selected features may lead us to the influential users.

4.5.1 Data Description

This experiment also uses the same tweets collection used in the experiment in section 4.3, which was a query on November 5th, 2013, with the words "باسم يوسف", from which we extracted a collection of 1221 unique users.

4.5.2 Method

We assume that each of the features selected as a result of experiment 4.2 reflects a trait presumed to be exhibited by influential users. The users are ranked according to each of the eight independent features and the top 50 users of each are included in separate lists. The users are then ranked according to their appearance frequency in these lists. We consider the users found at least once, then at least two times, three times, four times, five times and six times. The precision of each list will be calculated according to the manually assembled list of influential users which may be found in Appendix A.

4.5.3 Results

The users are sorted in descending order by their appearance frequency. The precision values are calculated for each top set as can be seen in Table 15. We look at the users found at

least once, then at least two times, three times, four times, five times and six times as may be seen in each of the columns, and the precision calculated.

Table 15: the Precision calculated for each

	Users found at least 1 time	Users found at least 2 times	Users found at least 3 times	Users found at least 4 times	Users found at least 5 times	Users found at least 6 times
Total number of users	224	102	47	21	6	2
Influential Users count	31	30	21	13	5	2
Precision	**0.14**	**0.29**	**0.45**	**0.62**	**0.83**	**1**

Figure 4 provides visual representation of the precision values in Table 15.

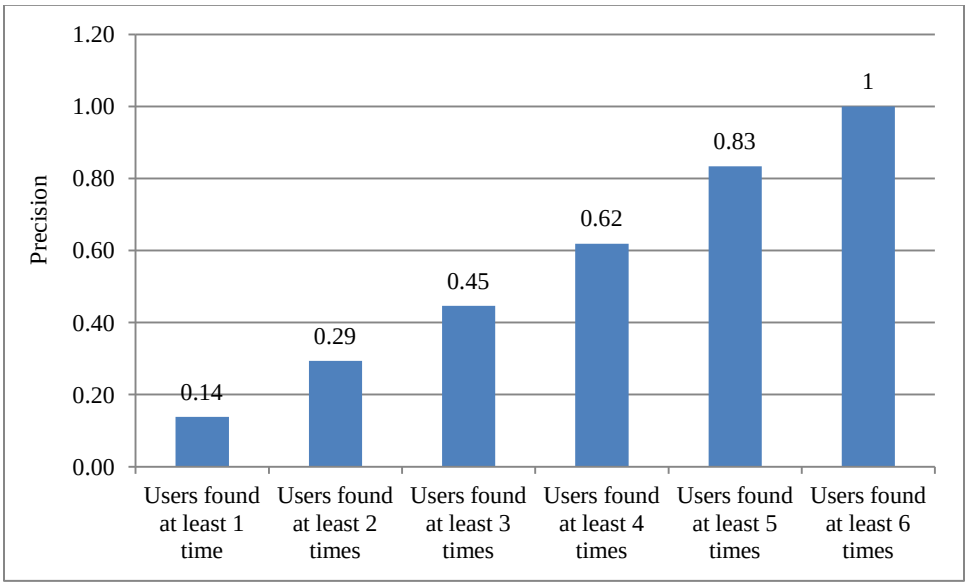

Figure 4: a visual of the precision values of experiment 4.5

4.5.4 Discussion

The users were ranked and extracted based on their appearance frequency in the rankings done according to each of the selected features. As we can see in Figure 4, the influential users' precision improved with the increase of the appearance frequency threshold. So if each of the features reflects a trait, then the more the traits a user exhibits the more likely they are influential.

4.6 Ranking Model Verification

The purpose of this experiment was to verify the effectiveness of the different ranking methods in detecting the influential users and to compare their performance.

4.6.1 Data Description

In January, February and March of 2014, twenty of the local trending topics on Google Trends and Twitter Trends were queried for this experiment. The queries are listed in Table 16, each with the number of tweets retrieved and the number of unique users specified.

Table 16: the topics queried

	Search Query	Query date	Number of retrieved tweets	Number of unique users
1	"السيسي"	26 Jan	2004	**1557**
2	"ميدان التحرير"	26 Jan	1738	**1332**
3	"تسلم الأيادي"	26 Jan	822	**732**
4	"عدلي منصور"	26 Jan	1550	**1206**
5	"مديرية امن القاهرة"	26 Jan	1424	**996**
6	"ترشيح السيسي"	28 Jan	1106	**971**
7	"مصر"	28 Jan	1900	**1342**
8	"25 يناير"	30 Jan	2003	**1262**
9	"30 يونيو"	30 Jan	2002	**1332**
10	"عنان"	29 Jan	2003	**1477**
11	"البلاوي"	26 Feb	2007	**1228**
12	"السيسي"	9 Feb	1618	**1195**
13	"باسم يوسف"	9 Feb	1936	**1449**
14	"سانت كاترين"	19 Feb	2005	**1521**
15	"سانت كاترين"	20 Feb	1623	**1237**
16	"طابا "	25 Feb	1179	**708**
17	"محلب"	25 Feb	2009	**1285**
18	"السيسي"	8 Mar	1810	**1224**
19	"محلب"	9 Mar	1760	**1061**
20	"مليون وحدة سكنية"	10Mar	1370	**1043**

4.6.2 Method

To each of the collections queried, the best of the ranking methods; those with the highest precision means, in sections 4.3, 4.4 and 4.5, are applied to rank the users. From experiment 4.3, we rank the users according to both the followers count and the average retweets count, from experiment 4.4, we rank the users according to both Score 3 and Score 4, and from experiment 4.5, we consider the users in the top 50 lists, having been ranked by each of the features independently, and appeared at least 5 times and at least 6 times. The top users of each are manually evaluated, abiding the same guide lines described in section 3.3.3, and their precision calculated. Finally, the T-test is conducted on some of the rankings' influential users' precision values to see if the difference or precision improvement is statistically significant.

4.6.3 Results

Table 17 is a summary of the precision values in this experiment. Each column represents one of the ranking methods used on the users, and each of the rows the queries. For each ranking carried out by each of the methods for each of the queries, we state the number of influential users found and the precision. For the rankings according to Followers count, Average Retweets, scores 3 and 4, the precision at 10 is calculated. As for the other two rankings, due to the presence of a tie, we do not cut at 10; we just focus on those who appeared at least 5 or 6 times regardless of their count. In case none of the users were found at least 6 times, we back-off to the users found at least 5 times. In the last row we calculate the average precision obtained by each of the ranking methods.

The T-test was conducted in order to compare between the different rankings carried out. The T-test null hypothesis was **not rejected** for any of the pairs of rankings' influential users' precisions.

Table 17: the influential users count and the precision values in experiment 4.6

	Search Query	Exp 4.3 - Followers count		Exp 4.3 - Average Retweets		Exp 4.4 - Score 3		Exp 4.4 - Score 4		Exp 4.5 - Users found at least 5 times		Exp 4.5 - Users found at least 6 times[1]	
		Influential Users count	Precision at 10	Influential Users count	Precision at 10	Influential Users count	Precision at 10	Influential Users count	Precision at 10	Influential Users count	Precision	Influential Users count	Precision
1	"السيسي"	10	**1.0**	8	**0.8**	8	**0.8**	10	**1.0**	4	**0.8**	2	**0.67**
2	"ميدان التحرير"	6	**0.6**	7	**0.7**	7	**0.7**	7	**0.7**	9	**0.75**	2	**1.0**
3	"تسلم الأيادي"	2	**0.2**	3	**0.3**	2	**0.2**	1	**0.1**	4	**0.4**	2	**1.0**
4	"عدلي منصور"	6	**0.6**	5	**0.5**	7	**0.7**	6	**0.6**	9	**0.53**	1	**0.33**
5	"مديرية امن القاهرة"	3	**0.3**	6	**0.6**	5	**0.5**	6	**0.6**	6	**0.3**	2	**0.5**
6	"ترشيح السيسي"	4	**0.4**	6	**0.6**	6	**0.6**	5	**0.5**	6	**0.5**	3	**0.75**
7	"مصر"	8	**0.8**	7	**0.7**	9	**0.9**	5	**0.5**	11	**0.92**	3	**1.0**
8	"25 يناير"	6	**0.6**	7	**0.7**	5	**0.5**	8	**0.8**	2	**0.67**	2	**0.67**[2]
9	"30 يونيو"	4	**0.4**	5	**0.5**	6	**0.6**	6	**0.6**	6	**1.0**	1	**1.0**
10	"عنان"	7	**0.7**	2	**0.2**	5	**0.5**	4	**0.4**	4	**0.67**	0	**0**[3]
11	"البلاوي"	8	**0.8**	7	**0.7**	7	**0.7**	8	**0.8**	8	**0.8**	1	**1.0**
12	"السيسي"	7	**0.7**	7	**0.7**	4	**0.4**	7	**0.7**	1	**0.5**	1	**0.5**[2]
13	"باسم يوسف"	6	**0.6**	5	**0.5**	5	**0.5**	6	**0.6**	1	**0.5**	1	**0.5**[2]
14	"سانت كاترين"	7	**0.7**	5	**0.5**	9	**0.9**	8	**0.8**	3	**1.0**	3	**1.0**[2]
15	"سانت كاترين"	9	**0.9**	7	**0.7**	10	**1.0**	10	**1.0**	4	**0.67**	4	**0.67**[2]
16	"طابا "	4	**0.4**	2	**0.2**	3	**0.3**	2	**0.2**	6	**0.35**	3	**0.5**
17	"محلب"	6	**0.6**	5	**0.5**	6	**0.6**	6	**0.6**	4	**0.36**	0	**0**[3]
18	"السيسي"	8	**0.8**	9	**0.9**	8	**0.8**	9	**0.9**	7	**0.88**	1	**1.0**
19	"محلب"	6	**0.6**	6	**0.6**	4	**0.4**	7	**0.7**	5	**0.7**	2	**1.0**
20	"مليون وحدة سكنية"	4	**0.4**	4	**0.4**	3	**0.3**	4	**0.4**	8	**0.57**	1	**0.5**
Precision means:		**0.605**		**0.565**		**0.595**		**0.625**		**0.644**		**0.68**	

[1] With back-off to 5 times when no users are found 6 or more times.
[2] Back-off is applied.
[3] Back-off was not applied, since none of the users found 6 or more times were considered by the manual evaluation to be influential.

Figure 5 reflects the values in Table 17, showing the precision values for each of the rankings carried out on each of the queries.

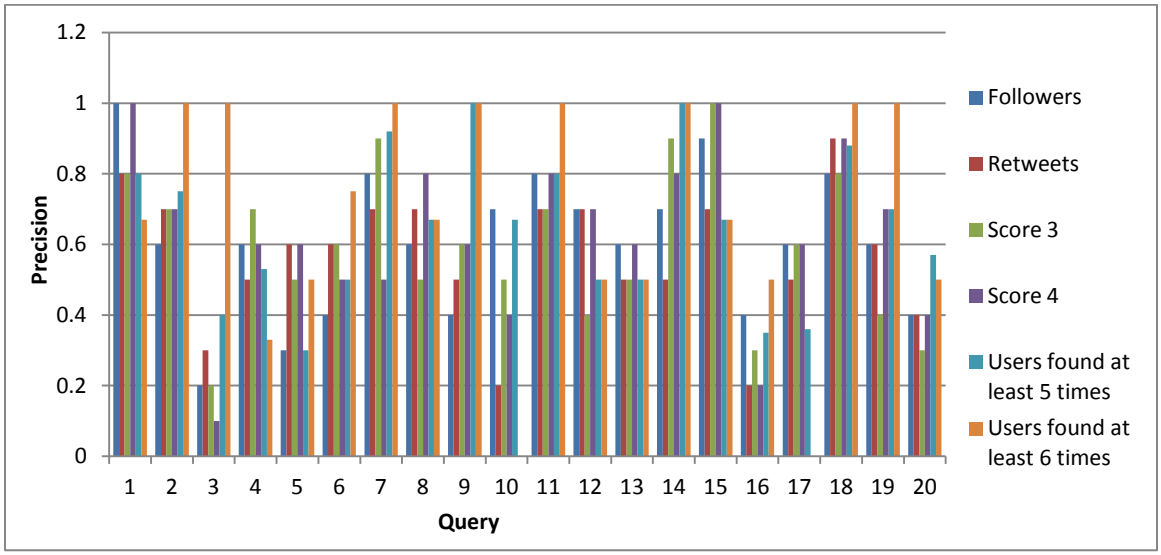

Figure 5: the precision values obtained for each of the queries by each of the rankings

4.6.4 Discussion

It should be noted that for Queries 8, 10, 12, 13, 14, 15 and 17, as seen in Table 17, the number of influential users count is zero for the set of users found at least 6 times. In the cases of Queries 10 and 17 none of the users found at least 6 times turned out to be influential, which resulted in the precision value of influential users to be zero. On the other hand, in case no users were found at least 6 times, which was the case with Queries 8, 12, 13, 14 and 15, we backed-off to the users found at least 5 times.

Each of the ranking methods was able to detect a set of influential users, however, their influential users' precisions varied from one query to another. As may be seen in Figure 5, there is no consistent outcome for any one of the ranking methods. As may be seen in Table 17, their precision means were relatively close. The highest precision mean obtained is 0.68 for the set of users found at least 6 times, where the precision of the influential users was 1 for eight out of the

20 queries. However, the differences between the different ranking outcomes were not found to be statistically significant.

From the investigations carried out for the evaluation process, we observed that the quality of attention a topic gets varies. Some topics maybe trending during a specific time and widely discussed and mentioned by many, but are not significant enough to attract the influential users or pull them into the discussion, which might explain the inconsistencies in the influential user presence from one topic to another. For example, the topics of Queries 1and 18 attracted a lot of influential users whereas the topic of Query 3 did not seem to get much attention from the influential users.

4.7 Summary

In experiment 4.1, we investigated the effectiveness of using SVM for account classification and compared it to the use of a manually prepared list of non-personal accounts. We found the SVM to be reliable and consistent with a precision of over 0.9. The results showed that over time the list performance deteriorates, and when the domain of the test data was changed, the SVM performed better than the lists, with higher precision and specificity values.

From a set of 12 relevant Twitter features, we selected 8 independent features in experiment 4.2, to be used in developing a model for detecting the influential users.

From ranking the users according to each of the selected features independently in experiment 4.3, we found that some of the features are better at detecting the influential users than others. The best four features were found to be Followers count, Average Retweets count, Average Retweet Frequency and the Account Age_Activity combination. These four features were used in experiment 4.4 as parameters in equations that would assign scores to each of the users. Score 3, which took the average of the Followers count, Average Retweets count and Average Retweet Frequency, and Score 4, which took the average of the Followers count and

Average Retweets count, gave better results than Score 1, which took the average of all four features, and Score 2, which took the average of Followers count, Average Retweets count and the Account Age_Activity combination feature.

In experiment 4.5 we investigated another approach for detecting influential users. Making use of the user rankings according to each of the eights selected features independently, as done in experiment 4.3, the users were ranked according to their appearance frequency in the lists. Using frequency thresholds to divide the users into sets, we found that the higher the frequency threshold the higher the influential users' precision is in the set. The best results were in the set of users found at least 5 times and at least 6 times in the top 50 users lists ranked according to each of the eight selected features; with the highest precision values of 0.83 and 1.0 respectively.

Finally in experiment 4.6, the best of the different ranking approaches from experiments 4.3, 4.4 and 4.5 were applied on the users of 20 different queried collections. Each of the ranking methods was able to detect influential users, however the precision varied between the queries. The highest precision mean, 0.68, was obtained from the set of users found at least 6 times. Eight of the 20 queries had a precision of 1.0, in compliance with the result of experiment 4.5 developing the method using just one collection of tweets. Also in compliance with the development experiment are the results of Score 4. In the development experiment, Score 4 produced the highest precision value, higher than Score 3 and any of the features when used independently, and in the verification experiment, Score 4 produced a higher precision mean value of 0.625; higher than that of Score 3, the Followers count and the Average Retweets count.

CHAPTER 5

Using Statistical Language Model for Detecting Influential Users

In this chapter we investigate the use of Statistical Language Modeling with the goal of finding out if the tweet language may be used as an indicator of influence. With the assumption that highly retweeted users are more likely to be influential, we test the use of Statistical Language Modeling to measure the quality of the tweet text and if it may be related to the retweets count of a tweet.

5.1 Using SLM: how the tweet perplexity relates to the other features

The objective of this experiment is to see how the perplexity values of the tweets relate to the users' features. Also to see the impact of increasing the training corpus size on the outcome of the statistical language model.

5.1.1 Data Description

For the training corpus of **Model 1**, 50 queries were carried out on a variety of popular topics during November and December of 2013, totaling to 71893 tweets. We filtered out the tweets with retweet count less than 20 and any duplicate texts. This brings the training corpus size of **Model 1** to **2354** tweets' text.

As for the training corpus of **Model 2**, 50 more queries were carried out during January of 2014, and added to the 50 queries of Model 1, totaling to 142050 tweets from 100 queries. We filtered out the tweets with retweet count less than 20 and any duplicate texts. This brings the training corpus size of **Model 2** to **4476** tweets' text.

All tweet texts are preprocessed; removing any non-Arabic text and symbols, before being used to train the language models.

As for the testing data, the same users' collection used in the experiment in section 4.2 to generate Table 9 was used in this experiment. There were a total of 10,539 tweets in the 6 queries posted by 5471 unique users. All 10,539 tweet texts were preprocessed the same as the training corpus; removing any non-Arabic text and symbols.

5.1.2 Method

Two statistical language models were generated. The difference between them is in the size of their training corpus size; **2354** tweet texts for training Model 1 and **4476** tweet texts for training Model 2.

To create a model from the training corpus, we first generate the n-gram count file from the training corpus, and then train the language model from the n-gram count file. Once the trained language models are ready, we use them to calculate the test data perplexity. The perplexity value for each of the 10,539 test tweets was calculated by each of the models (Model 1 and Model 2). Since a user may have more than one tweet, for each of the 5471 users we calculated their tweets' average perplexity values. To see how the users' average perplexity values of the tweets relate to the users' other features; the features selected in experiment 4.2, we calculate the correlation values between each of the eight selected users' features and their average tweets' perplexity values.

5.1.3 Results

Tables 18 and 19 contain the correlation values between the users' average perplexity values and their other features. Table 18 uses the perplexity values calculated using Model 1, and Table 19 uses the perplexity values calculated using Model 2.

Table 18: the correlation values between the users' features and their average tweets' perplexity values calculated by Model 1

	Avg. Activity Rate	Age + Activity Combo	Followers	TFF Ratio	Coll. Tweets Count	Avg. Retweet Count	Avg. Tweet Age	Retweet Freq.
Avg Perplexity values	**0.061**	**-0.001**	**-0.031**	**0.009**	**0.035**	**-0.070**	**-0.009**	**0.010**

Table 19: the correlation values between the users' features and their average tweets' perplexity values calculated by Model 2

	Avg. Activity Rate	Age + Activity Combo	Followers	TFF Ratio	Coll. Tweets Count	Avg. Retweet Count	Avg. Tweet Age	Retweet Freq.
Avg Perplexity values	**0.049**	**0.007**	**-0.029**	**0.008**	**0.035**	**-0.067**	**-0.004**	**0.011**

5.1.4 Discussion

From the numbers in Table 18 and 19, there is an inverse correlation between the tweets' average perplexity values and some of the features, the lowest values are those of the correlation between the average retweets count and the average perplexity values at -0.070 in Table 18, and -0.067 in Table 19. Since the training corpus used to estimate the Statistical Language Model was composed of the highly retweeted tweets, it shows that the best correlation being between the perplexity and the retweets is a support to our hypothesis; that the SLM can be trained to detect popular tweets. There is also an inverse correlation between the followers counts and the average perplexity values. The assumption is that since there exists a correlation of 0.478, as can be seen in Table 9, between the followers counts and the average retweets counts, then an inverse correlation, close to that of the average retweets counts, is also bound to exist.

Increasing the size of the training corpus didn't have much of an effect on the outcome of the model. The difference in the correlation values between the features and the perplexity values, as a result of Model 1 and Model 2, was very small. As previously mentioned, the most frequent

terms in one hour or day tend to be very different from those in the next, significantly more so on Twitter than in other content on the web. 17% of the top 1000 query terms "churn over" on an hourly basis. During major events, the frequency of queries spikes dramatically (Twitter, 2012c). This rapid change makes the language model estimation, which relies on term or phrase frequencies, more challenging, which explains why increasing the size of the training, especially at a later time, did not have the expected outcome on the language model.

5.2 Using SLM: perplexity and average tweet word count

The objective of this experiment is to see if there is a relation between the users' tweets average word count and perplexity values and to find out the average tweet word count tendency of the influential users.

5.2.1 Data Description

This experiment also uses the same tweets collection used in the experiment in section 4.3, which was a query on Novermber 5th, 2013, with the words "باسم يوسف", from which we extracted a collection of 1221 unique users. All 1593 of the collection tweets texts were preprocessed; removing any non-Arabic text and symbols. For the influential users in the collection, we refer to the list of users in Appendix A.

5.2.2 Method

First, a graph of the users' average tweets word counts is plotted against the average tweet perplexity values, to visualize the relationship between the two, and their correlation calculated. Then we analyze the average word count of the influential users.

5.2.3 Results

In Figure 6, for all 1221 users, the average tweets word counts is plotted against the average tweet perplexity values. The influential users' points are highlighted.

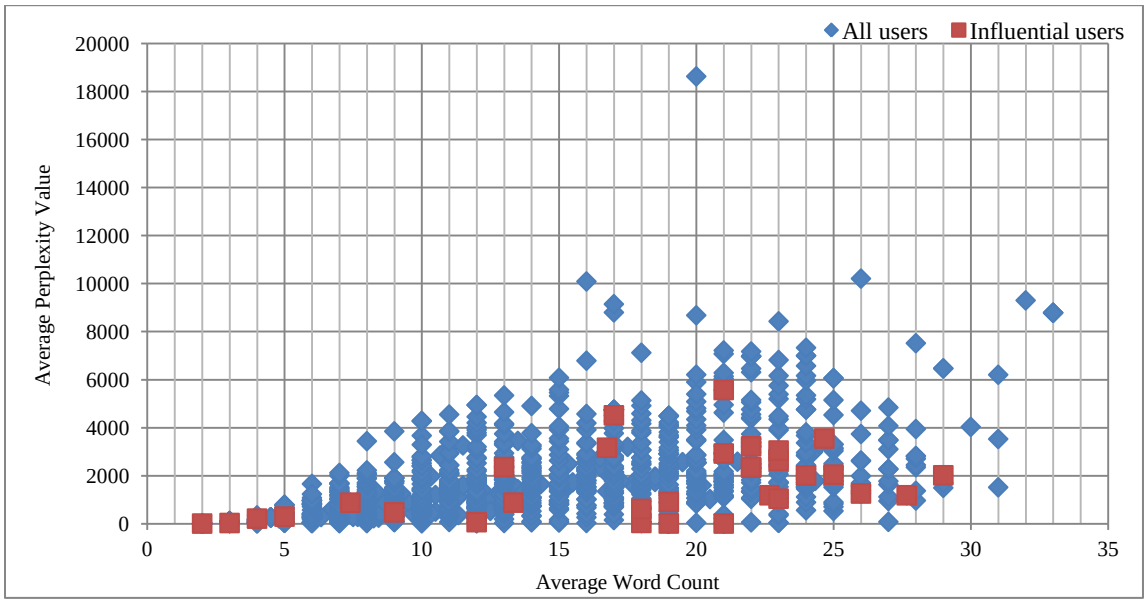

Figure 6: the users' tweets' average perplexity values plotted against the users' average word count

From the average perplexity values and the average word count pairs used to plot the graph in Figure 6, they were found to have a correlation of **0.52.** Further analysis of the points in Figure 6 showed that the average word count range was [2 – 33] with a mean of **14.14**. As for the influential users, their average word count range was [2-29] with a mean of **17.85**.

The average word counts were divided into three ranges and the number of influential users within each range counted, as may be seen in Figure 7.

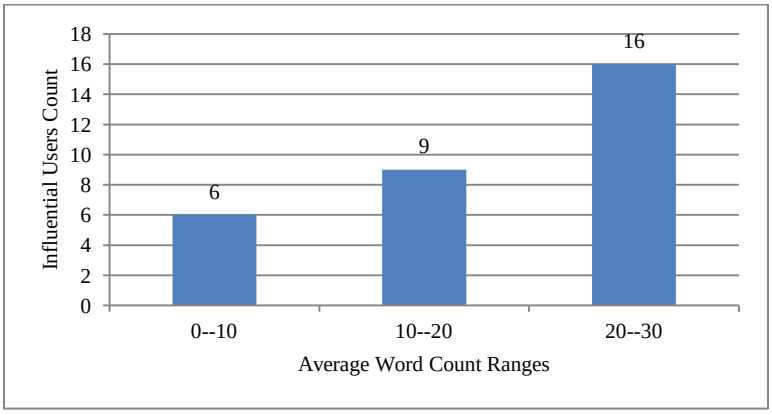

Figure 7: the influential users' distribution within the average word count ranges

5.2.4 Discussion

There is a correlation between the average word counts and the average perplexity values of 0.52, which is visible in Figure 6. The longer the sentence, the higher the perplexity value may be. We also found out that most influential users tend to write longer sentences. That may be deduced from their distribution within the word count ranges in Figure 7, and their average word count mean of 17,85.

5.3 Using SLM: the perplexity values as a feature for detecting influential users

The objective of this experiment is to see the effect of ranking the users according to their perplexity values. We want to find out if the perplexity values may be used as supporting features for detecting topic-specific influential users.

5.3.1 Data Description

This experiment uses the same tweets collection used in the experiment in section 4.3, which was a query on November 5th, 2013, with the words "باسم يوسف", from which we extracted a collection of 1221 unique users. All 1593 of the collection tweets texts were preprocessed; removing any non-Arabic text and symbols.

5.3.2 Method

We measure the perplexity values of all the tweet texts using Model 1, from the experiment in section 5.1. Each user is associated with the average perplexity value and the minimum perplexity value of the tweets they posted.

The users are ranked in ascending order, once according to the average perplexity values, and again according to the minimum perplexity value. The effectiveness of each of the rankings is evaluated according to the manually assembled list of influential users which may be found in Appendix A.

5.3.3 Results

The users are ranked twice; once according to their average perplexity values and another according to their minimum perplexity values, where the top 50 users may be seen in Tables 13 and 14, respectively, in Appendix B. The precision at 10, 20, 30, 40 and 50 for each of the rankings are calculated as may be seen in the columns in Table 20. In the rows are the influential users count and the calculated precision for each of the two rankings.

Table 20: the influential users count and the precision values in experiment 4.8

		Top 10	Top 20	Top 30	Top 40	Top 50
ranked according to the **average** perplexity	Influential Users Count	1	1	2	2	2
	Precision	0.1	0.05	0.067	0.05	0.04
ranked according to the **minimum** perplexity	Influential Users Count	0	0	2	7	9
	Precision	0	0	0.07	0.175	0.18

Figure 8 provides visual representation of the precision values in Table 20.

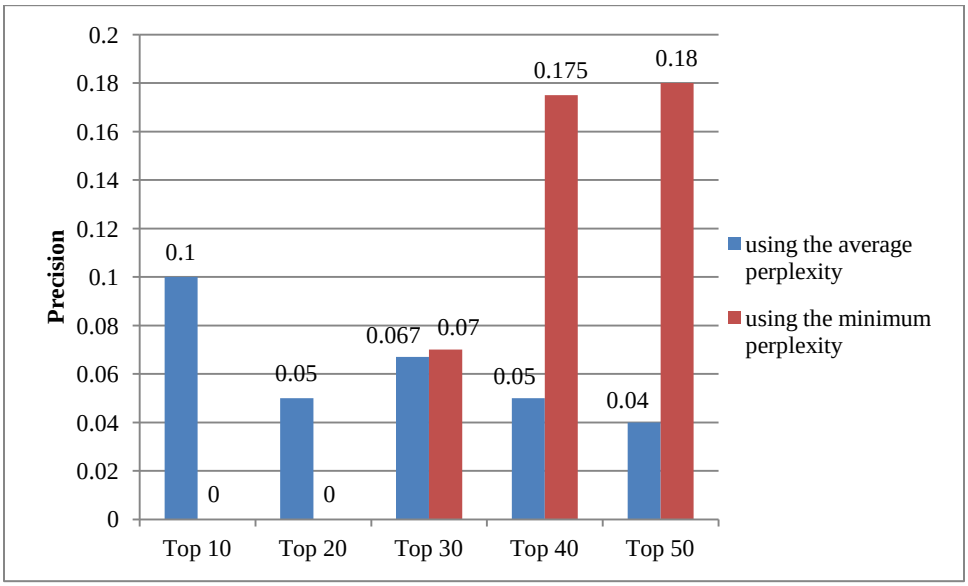

Figure 8: a visual of the precision values of experiment 4.8

5.3.4 Discussion

Ranking the users just according to their tweets' perplexity values resulted in very low precision values for influential users, as may be seen in Figure 8. Also after a closer look at the top users ranked according to perplexity, in Tables 13 and 14 in Appendix B, we found that

several users had the same perplexity value of 9.377. After referring back to the tweet texts to investigate the reason for that, we found that what the users had in common was that their tweets consisted of a hashtag "#باسم يوسف" and a web-link, and since our tweet text preprocessing removes all non-Arabic text and symbols, the text that underwent evaluation by the model was the same, which would explain the similar perplexity value for all these users. As for the reason why that particular phrase or name got the least perplexity value, the Model 1 counts file showed that the highest n-gram frequencies were those including the bi-gram "باسم يوسف". This shows that the perplexity cannot be used for indication of influence.

5.4 Using SLM: incorporating the perplexity in the user ranking

The objective of this experiment is to incorporate the perplexity values in the ranking method. We want to see the effect of using perplexity in the ranking process and the effect of filtering out users with lower average word counts.

5.4.1 Data Description

This experiment also uses the same tweets collection used in the experiment in section 4.3, which was a query on Novermber 5th, 2013, with the words "باسم يوسف", from which we extracted a collection of 1221 unique users.

5.4.2 Method

We found that ranking this collection's users according to score 4 of experiment 4.4 produced a high precision at 10 for influential users of 0.9. So using one of the ranking methods we experimented with, we try incorporating the user's perplexity value in the ranking. So for this experiment we first rank the users according to Score 4 of experiment 4.4.

$$Score\ 4 = \frac{1}{2}(Normalized\ \boldsymbol{F} + Normalized\ \boldsymbol{RT})$$

The top 50 users are then re-ranked, in ascending order, six times: once according to their average perplexity values and another according to their minimum perplexity values. Users with average word count less than 10 are then filtered out and the users are re-ranked again, once according to their average perplexity values and another according to their minimum perplexity values. Then finally, users with average word count less than 20 are then filtered out and the users are re-ranked yet again, once according to their average perplexity values and another according to their minimum perplexity values.

5.4.3 Results

The ranking of influential users' precision at 10 values of the different re-rankings may be seen in Table 21, where the column specifies the perplexity values used to rank, and the rows specify the word count threshold applied.

Table 21: the precision values of the re-rankings

	Using the **Average** Perplexity	Using the M**inimum** Perplexity
no word count threshold	0.2	0.6
word count >=10	0.4	0.6
word count >=20	0.4	0.4

Figure 9 provides visual representation of the precision values in Table 21.

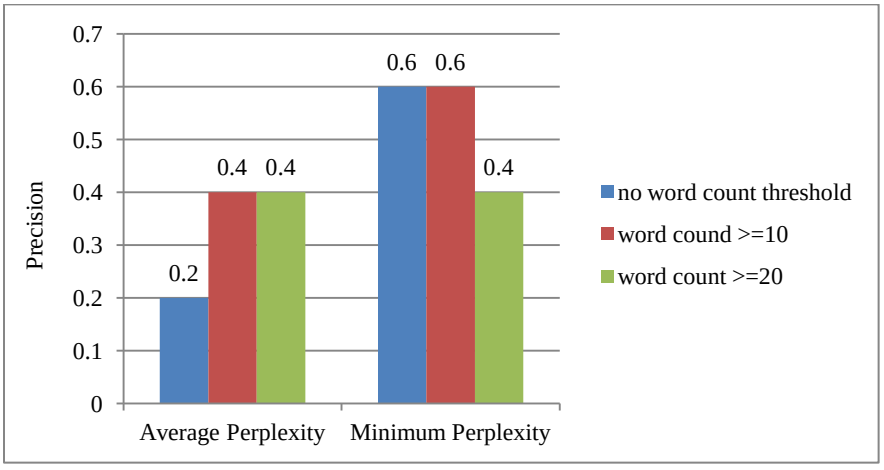

Figure 9: the precision at 10 values for each of the re-rankings

5.4.4 Discussion

The results show that ranking using the minimum perplexity values results in higher influential user precision values. Also, putting a word count threshold improved the ranking with the average perplexity values.

The highest precision obtained in this experiment was 0.6. This precision is not an improvement over the original 0.9 obtained by Score 4 in experiment 4.4. These results show that, incorporating the perplexity into the ranking method did not improve the influential users' precision at 10.

5.5 Using SLM: Verification

The objective of this experiment is to verify the performance of experiment 5.4 and see how the results vary with different data.

5.5.1 Data Description

This experiment uses the same data used in experiment 4.6; the same 20 topics, listed in Table 16.

5.5.2 Method

We repeat the same steps carried out in experiment 5.4 on each of the queries listed in Table 16. We first rank the users according to Score 4 of experiment 4.4. The top 50 users are then re-ranked, in ascending order, six times; once according to their average perplexity values and another according to their minimum perplexity values. Users with average word count less than 10 are then filtered out and the users are re-ranked again, once according to their average perplexity values and another according to their minimum perplexity values. Then finally, users with average word count less than 20 are then filtered out and the users are re-ranked yet again, once according to their average perplexity values and another according to their minimum

perplexity values. Finally, the T-test is conducted on the rankings' influential users' precision values to see if the differences are statistically significant.

5.5.3 Results

Table 22 shows the precision at 10 values for each of the six different rankings carried out on each of the 20 queries. The different ranking approaches may be seen in the columns, and each of the queries in a row. The final row shows the mean precision obtained by each of the rankings.

The T-test was conducted in order to compare between the different rankings carried out. The T-test null hypothesis was **not rejected** for any of the pairs of rankings' influential users' precisions.

Table 22: the influential users count and the precision values in experiment 5.5

	Search Query	No word count threshold		word count >=10		word count >=20	
		Using Average Perplexity	Using Minimum Perplexity	Using Average Perplexity	Using Minimum Perplexity	Using Average Perplexity	Using Minimum Perplexity
1	"السيسي"	**0.7**	**0.6**	**0.8**	**0.8**	**0.6**	**0.5**
2	"ميدان التحرير"	**0.3**	**0.3**	**0.5**	**0.6**	**0.6**	**0.6**
3	"تسلم الأيادي"	**0.2**	**0.2**	**0.3**	**0.3**	**0.2**	**0.2**
4	"عدلي منصور"	**0.2**	**0.2**	**0.3**	**0.2**	**0.5**	**0.5**
5	"مديرية امن القاهرة"	**0**	**0.2**	**0**	**0.2**	**0.2**	**0.2**
6	"ترشيح السيسي"	**0.6**	**0.6**	**0.5**	**0.5**	**0.6**	**0.6**
7	"مصر"	**0.6**	**0.6**	**0.6**	**0.7**	**0.9**	**0.8**
8	"25 يناير"	**0.8**	**0.8**	**0.8**	**0.8**	**0.5**	**0.4**
9	"30 يونيو"	**0.2**	**0.3**	**0.4**	**0.4**	**0.4**	**0.4**
10	"عنان"	**0.6**	**0.6**	**0.5**	**0.6**	**0.4**	**0.4**
11	"البلاوي"	**0.3**	**0.3**	**0.1**	**0.2**	**0.2**	**0.3**
12	"السيسي"	**0.4**	**0.4**	**0.4**	**0.6**	**0.5**	**0.8**
13	"باسم يوسف"	**0.4**	**0.4**	**0.5**	**0.6**	**0.5**	**0.5**
14	"سانت كاترين"	**0.7**	**0.6**	**0.7**	**0.6**	**0.8**	**0.8**
15	"سانت كاترين"	**0.8**	**0.6**	**0.8**	**0.7**	**1.0**	**0.9**
16	"طابا "	**0.1**	**0.1**	**0.1**	**0.1**	**0.2**	**0.2**
17	"محلب"	**0.3**	**0.2**	**0.3**	**0.2**	**0.3**	**0.2**
18	"السيسي"	**0.5**	**0.6**	**0.5**	**0.5**	**0.5**	**0.5**
19	"محلب"	**0.3**	**0.3**	**0.3**	**0.3**	**0.4**	**0.4**
20	"مليون وحدة سكنية"	**0**	**0**	**0**	**0**	**0.2**	**0.2**
Precision mean:		**0.4**	**0.395**	**0.42**	**0.445**	**0.475**	**0.47**

Figure 10 shows the precision of the influential users for each of the six re-rankings on each of the 20 queries.

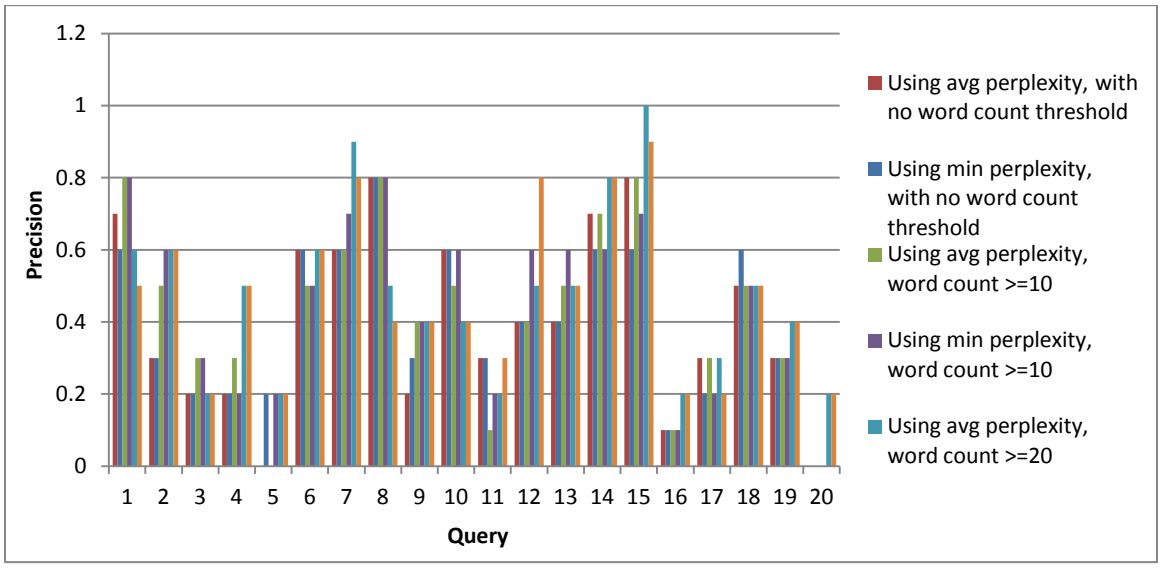

Figure 10: the precision at 10 values for each of the queries

5.5.4 Discussion

Each of the re-rankings was able to detect a set of influential users, however, their precisions varied from one query to another. As may be seen in Figure 5, there is no consistent outcome for any one of the ranking methods. The precision values are affected by the queried collection itself, consolidating the observation that the quality of attention a topic attracts varies.

As may be seen in Table 22, in a few cases, such as with Queries 3, 6, 7 and 10, the re-rankings resulted in higher influential users' precision values than the initial ranking done according to Score 4 (the values of which may be seen in Table 17 of experiment 4.6). However, the precision mean obtained by ranking users according to Score 4 is 0.62, as may be seen in Table 17, which is higher than any of the precision mean values obtained by any of the re-rankings, as may be seen in Table 22.

The precision mean values suggest that the use of perplexity in the ranking process does not improve the detection of influential users in most of the cases.

5.6 Summary

In experiment 5.1 we see how the perplexity values of the users' tweets relate to their other features. We found that since the training corpus of the SLM consisted of the highly retweeted tweet texts, the least inverse correlation was between the users' average perplexity values and the retweet counts. This consolidates the hypothesis that the SLM can be trained to detect the highly retweeted tweets. As for increasing the training corpus size, it did not have much of an effect on the outcome of the model.

In experiment 5.2 we found a high correlation of 0.52 between the users' average perplexity values and their average tweet word count. We also found that most influential users tend to write longer sentences with higher word counts.

When the users were ranked solely according to their perplexity values in experiment 5.3, the rankings were found to have very low precision for the influential users. So in experiment 5.4 we investigated incorporating the users' perplexity values in Score 4; one of the ranking methods of chapter 4 that resulted in high precision for the influential users. We also investigated the effect of putting thresholds on the users' word count; filtering out users below the threshold. We found that ranking users according to their minimum perplexity values resulted in higher influential users precision than when ranking according to the users' average perplexity values. However, the results showed that incorporating the use of perplexity values and word count thresholds did not improve upon the original ranking done according to Score 4. So in order to verify the performance of this experiment, in experiment 5.5 we repeated the experiment, carrying out the same steps on the users of 20 different sets of queried tweets collections. Each of the rankings was able to detect a set of influential users, however, their precision varied with no consistency found in the outcomes for any one of the ranking methods. Also, despite the presence of a few cases where the use of perplexity and word count threshold did improve upon the

original ranking, the overall precision mean values still suggest that the original ranking was better.

CHAPTER 6

Conclusion and Future Work

We are interested in the problem of identifying which Twitter users may have influence on fellow users in a specific topic. The micro-blogging service Twitter has become a very popular tool for expressing opinions, broadcasting news, and simply communicating with friends and people. Twitter is not only interesting because of its real-time response, but also because it is sometimes ahead of the newswire.

Much analysis on the data available by the API has been done and there has been a broad spectrum of approaches proposed. We found multiple approaches for detecting influential members is social networks. Social Network Analysis was used in multiple researches to measure the relationships between network members. There are many key figures which describe the position and communication habits of users to analyze the interaction network in order to find the influential users, the most popular of which is the centrality analysis. There was another approach using a modified K-shell decomposition algorithm. As interesting and compelling as these studies seemed, we decided not to go through a similar approach when addressing our problem. There were several other approaches that rely on mathematical models and/or algorithms to quantify user influence on social networks using a set of intuitive properties that can be approximated by some collectable statistics. Also some linguistic analysis approaches were investigated. We were inspired by such approaches when addressing our problem.

The objective of this research is to detect the influential users in a specific topic on Twitter. In more detail, from a collection of tweets matching a specified query, we want to detect the influential users, in an online fashion. In order to address this objective, we first want to focus our search on the individuals who write in their personal accounts, so we investigate how we can differentiate between the personal and non-personal accounts. Secondly, we investigate which set of features can best lead us to the topic-specific influential users, and how these features can be

expressed in a model to produce a ranked list of influential users. Finally, we look into the use of the language and if it can be used as a supporting feature for detecting the author's influence.

To address the problem of detecting the influential Twitter users we developed a data collection tool to retrieve the necessary data from Twitter. Firstly, since we only want to include the personal accounts, we carried out account classification using SVM and compared that to using a manually assembled list of the mom-personal accounts. Then having determined the relevant features and had them tested for intra-dependencies, user ranking methods were developed and evaluated. Finally, the use of a statistical language model (SLM) for tweet text evaluation was investigated to see if the user's language may also be used as an influence indicator.

For user account classification, the performance of both the SVM and manually assembled list were pretty close in some cases, however, our results showed that the use of SVM is more reliable since it is domain independent and should not decay with time as the manually assembled list does. The results also show that account classification, using a set of basic account features, such as the followers, friends, listed, statuses count and the average daily activity rate; instead of analyzing temporal patterns and users' past behavior, produces good results with a precision values over 0.9. The results showed that the manually assembled list performance deteriorates over time and that when the domain of the test data is changed, the SVM performed better than the lists, with higher precision and specificity values.

In order to decide on which from the set of relevant features to use, the correlation values between each of the features were calculated. A high correlation implies dependency between the features, and when there is a high dependency between two features, using both is redundant. Having settled on a set of eight independent features, we relied on these features in the experiments to develop the model that would detect the influential users.

The lack of an obvious reference point regarding influential users on Twitter and the inapplicability of the evaluation approaches we reviewed in the literature, led us to resort to a labor intensive manual evaluation approach. In order to produce calculable measures, we studied the users; their tweets and profile pages, and manually decided on who the relevant influential users are, and according to which precision values were calculated.

From ranking the users according to each of the eight selected features independently, we found that the Followers count, Average Retweets count, Average Retweet Frequency, and the Age_Activity combination features were the best at ranking the influential users at the top. Two ranking methods were developed to combine these best four features. In the first method, we combined the best four features into equations and the users were ranked according to the resulting score. This method was able to obtain high precision at 10 values of up to 0.8 and 0.9 for the equations of Score 3, which took the average of the Followers count, Average Retweets count and Average Retweet Frequency, and Score 4, which took the average of the Followers count and Average Retweets count. In the second method, the users ranked in the top 50 according to each of the eight selected independent features were divided into sets according to their appearance frequency in the lists. This method was able to obtain the highest precision values of up to 0.83 and 1.0 for the sets of users found at least 5 times and those found at least 6 times respectively in the eight lists. Both ranking methods were then conducted on 20 queries to verify their effectiveness in detecting influential users, and compare their performance. The set of users found at least 6 times (in the top 50 ranked according to each of the eight selected features) was found to have the most consistent outcome and the highest precision mean of 0.692.

With the objective of capturing a quality exhibited by highly retweeted content, we investigated the use of statistical language analysis. Using a large collection of highly retweeted tweet texts as a training corpus, a statistical language model was estimated. Several collections were evaluated by the model and tested to determine if the tweet text perplexity value can be used

as a linguistic feature. An inverse correlation was found between the users' tweets' average perplexity values and the average retweets count. This supports the hypothesis that the SLM can be trained to detect highly retweeted posts. However, when the perplexity was used in ranking the users, the precision of influential users was very low. The nature of the language and people's writing style on Twitter is all too diverse to be comprehensively captured by a language model. The twitter community in general is very tolerable of the improper use of the language which has become quite common as of late; bad grammar or lake thereof, and flexible spelling and abbreviations. That in addition to the Arabic dialect often used, which is unbound by any rules and varies across different regions and/or communities.

The contributions of this thesis can be summarized into the following. A method of classifying accounts as personal or non-personal was proposed. The features that help detecting influential users were identified to be the Followers count, the Average Retweets count, the Average Retweet Frequency and the Age_Activity combination. Two methods for identifying the influential users were proposed. Finally, the simplistic approach using SLM did not produce good results, and there is still a lot of work to be done for the SLM to be used for identifying influential users.

For future work investigation ideas we propose exploring other API options and consider new features, including opinion polarity and study its effect on influence, finding a quantifiable measure for eloquence and/or readability in tweets, and studying different preprocessing s to improve the textual language model outcome.

References

N. Agarwal, H. Liu, L. Tang, and P. S. Yu, “Identifying the influential bloggers in a community,” *Proceedings of the international conference on Web search and web data mining - WSDM ’08*, p. 207–217, 2008.

L. Akritidis, D. Katsaros, and P. Bozanis, “Identifying Influential Bloggers: Time Does Matter,” *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 76–83, 2009.

L. Akritidis, D. Katsaros, and P. Bozanis, “Identifying the Productive and Influential Bloggers in a Community,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 759–764, Sep. 2011.

S. Alshaer and F. Salem, “The Arab World Online: Trends in Internet Usage in the Arab Region”, a White Paper produced by Dubai School of Government’s Governance and Innovation Program, April 2013

I. Anger and C. Kittl, “Measuring Influence on Twitter,” *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '11*, September 2011.

R. Baeza-Yates and B. Ribeiro-Neto, “Modern information retrieval,” 1999.

L. R. Bahl, J. K. Baker, F. Jelinek and R. L. Mercer, “Perplexity - a measure of the difficulty of speech recognition tasks” Program of the 94th Meeting of the Acoustical Society of America J. Acoust. Soc.Am., 62:S63, 1977

E. Bakshy, J. M. Hofman, D. J. Watts, and W. A. Mason, “Everyone’s an Influencer : Quantifying Influence on Twitter,” *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, pp. 65–74, 2011.

A. Baron, V. Punyakanok and M. Freedman, "Using Signals from Text to Identify Roles within a Group," *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on* , vol., no., pp.38,44, 19-21 Sept. 2012

C. Bigonha and T. N. C. Cardoso, “Detecting evangelists and detractors on twitter,” *Proceedings of the Brazilian Symposium on Multimedia and the Web, pp. 107–114, 2010.*

F. Bodendorf and C. Kaiser, “Detecting opinion leaders and trends in online social networks,” *Proceeding of the 2nd ACM workshop on Social web search and mining - SWSM ’09*, p. 65–*68*, 2009.

L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees,” *Chapman and Hall/CRC*, 1984.

P. E. Brown and J. Feng, "Measuring User Influence on Twitter Using Modified K-Shell Decomposition," *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 18–23, 2011.

J. del Campo-Ávila, N. Moreno-Vergara and M. Trella-López, "Analizying Factors to Increase the Influence of a Twitter User ," *Highlights in Practical Applications of Agents and Multiagent Systems Advances in Intelligent and Soft Computing* Volume 89, pp 69-76, 2011.

S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "New Model of Internet Topology Using k-shell Decomposition," *Proceedings of the National Academy of Sciences 104.27*, pp. 1–5, 2006.

M. Cha, H. Hamed, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter : The Million Follower Fallacy," *Fourth International AAAI Conference on Weblogs and Social Media, pp. 10–17*, 2010.

C. Chang and C. Lin, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

B. Chen, "Course notes for CS114: Introduction to SRILM Toolkit", Available: http://www.cs.brandeis.edu/~cs114/CS114_docs/SRILM_Tutorial_20080512.pdf, May 2008, viewed on August 25, 13.

M. Collins, "Course Notes for COMS w4705: Language Modeling", Available: http://www.cs.columbia.edu/~cs4705/fall2012/notes/lm.pdf, 2012, viewed on August 23, 13

Y. Cui, X. Chen, and X. Li, "The topology analyze of blogosphere through social network method," *2011 Seventh International Conference on Natural Computation*, pp. 302–305, Jul. 2011.

C. Danescu-Niculescu-Mizil, M. Gamon and S. Dumais, "Mark my words!: linguistic style accommodation in social media," *Proceedings of the 20th international conference on World Wide Web - WWW '11*, March 2011.

D. Donaldson and D. Hounshell, "TFF Ratio", Available: http://tffratio.com/Default.aspx, 2009, viewed on December 7, 2013.

D. W. Drezner and H. Farrell, "The power and politics of blogs," *American Political Science Association*, 2004.

Z. Fox, "How the Arab World Uses Facebook and Twitter", Available: http://mashable.com/2012/06/08/arab-world-facebook-twitter/, 8 June 2012, viewed on September 8, 2013.

D. Gaffney and C. Puschmann, "Game or measurement? Algorithmic transparency and the Klout score," Symposium and Workshop on Measuring Influence on Social Media (pp. 1–2), 2012.

M. A. Graber, C. M. Roller and B. Kaeble, "Readability of Patient Education Material on the World Wide Web", *Journal of Family Practice 48(1)*, 58, 1999.

T. H. Haveliwala, "Topic-sensitive PageRank," *Proceedings of the eleventh international conference on World Wide Web - WWW '02*, p. 517–526, 2002.

E. Keller and J. Berry, "One American in ten tells the other nine how to vote, where to eat, and what to buy," *They are The Influentials, New York*, vol. 25, no. 5, pp. 1–8, 2003.

E. Kiciman, "Language differences and metadata features on Twitter," *Web N-gram Workshop*, 2010.

M. Kitsak, L. K. Gallos, S. Havlin, F. Liljerros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics 6.11*, pp. 888–893, 2010.

N. Kokash, "An introduction to heuristic algorithms,"2005.

M. McPherson, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, 2001.

F. Meng, J. Wei, and Q. Zhu, "Study on the Impacts of Opinion Leader in Online Consuming Decision," *2011 International Joint Conference on Service Sciences*, pp. 140–144, May 2011.

Microsoft, "CORREL" [Microsoft Office Support] Available: http://office.microsoft.com/en-001/excel-help/correl-HP005209023.aspx?CTT=1, 2014, viewed on March 30, 14.

D. Murthy, A. Gross, and D. Oliveira, "Understanding Cancer-Based Networks in Twitter Using Social Network Analysis," *2011 IEEE Fifth International Conference on Semantic Computing*, pp. 559–566, Sep. 2011.

M. Naaman, J. Boase and C. Lai, "Is it really about me?: Message Content in Social Awareness Streams, "*Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, February 2010.

V. Pareto, A. N. Page, "Translation of *Manuale di economiapolitica ("Manual of political economy")"*, A.M. Kelley, 1971.

H. Purohit, Y. Ruan, A. Joshi, S. Parthasarathy and A. Sheth, "Understanding User-Community Engagement by Multi-faceted Features: A Case Study on Twitter", in SoME 2011 Workshop on Social Media Engagement, in conjunction with WWW 2011, March 2011

D. Quercia, J. Ellis, L. Capra, and J. Crowcroft, "In the Mood for Being Influential on Twitter," *2011 IEEE third international conference on security, risk and trust (passat), and 2011 IEEE third international conference on social computing (socialcom)*, pp. 307–314, 2011.

A. Rafea et. al., "Initial SATA Prototype Implementation", Sentiment Analysis and Opinion Mining of the Arabic Web (Digital Content), www.cse.aucegypt.edu/~rafea/SATA, 2012

D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and Passivity in Social Media,"*Proceedings of the 20th international conference companion on World wide web - WWW '11*, 2010.

R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where do we go from here?, " *Proceedings of IEEE, vol.88, no.8, pp.1270-1278,* August 2000.

M. Shalaby and A. Rafea**, "**Identifying the Topic-Specific Influential Users and Opinion Leaders in Twitter ", *Proceedings of Artificial Intelligence and Applications -AIA 2013*, February 2013.

A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," *Proceedings of the* International Conference *on* Spoken Language Processing, September 2002.

A. Stolcke, J. Zheng, W. Wang and V. Abrash, "SRILM at Sixteen: Update and Outlook," *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, December 2011.

B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network," *2010 IEEE Second International Conference on Social Computing*, pp. 177–184, Aug. 2010.

W. Sun and H. Qiu, "A social network analysis on Blogospheres," *2008 International Conference on Management Science and Engineering 15th Annual Conference Proceedings*, pp. 1769–1773, Sep. 2008.

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

G.Tavares and A. Faisal, "Scaling-Laws of Human Broadcast Communication Enable Distinction between Human, Corporate and Robot Twitter Users," *PLoS ONE* 8(7): e65774, 2013.

Twitter, "Celebrating #Twitter7" [Twitter Blog] Available: https://blog.twitter.com/2013/celebrating-twitter7, 2013, viewed on October 6, 13.

Twitter, "Twitter, in your language" [Twitter Blog] Available: http://blog.twitter.com/2012/01/twitter-in-your-language.html, 2012a, viewed on August 7, 12

Twitter, "Twitter Turns Six" [Twitter Blog] Available:http://blog.twitter.com/2012/03/twitter-turns-six.html, 2012b, viewed on August 7, 12.

Twitter, "Studying rapidly evolving user interests" [Twitter blog] Available: https://blog.twitter.com/2012/studying-rapidly-evolving-user-interests, 2012c, viewed on September 10, 13.

Twitter, "200 Million Tweets per day" [Twitter Blog] Available: http://blog.twitter.com/2011/06/200-million-tweets-per-day.html, 2011, viewed on August 7, 12.

Twitter API Documentation, “Platform Objects: Users” Available: https://dev.twitter.com/docs/platform-objects/users, 2013a, viewed on August 24, 13

Twitter API Documentation, “Platform Objects: Tweets” Available: https://dev.twitter.com/docs/platform-objects/tweets, 2013b, viewed on August 24, 13

E. Vega, R. Parthasarathy and J. Torres, “Where are my Tweeps?: Twitter Usage at Conferences,” *Paper, Personal Information*, pp. 1–6, 2010.

L. Wang, T. Lou, J. Tang, and J. E. Hopcroft, “Detecting Community Kernels in Large Social Networks,” *2011 IEEE 11th International Conference on Data Mining*, pp. 784–793, Dec. 2011.

W. Webberley, S. Allen, and R. Whitaker, “Retweeting: A study of message-forwarding in twitter,” *2011 Workshop on. IEEE Mobile and Online Social Networks (MOSN)*, pp. 13–18, 2011.

J. Weng, E. Lim, J. Jiang, and Q. He, “TwitterRank : Finding Topic-sensitive Influential Twitterers,” *Proceedings of the third ACM international conference on Web search and data mining*, 2010.

Wikipedia, “The Hirsch H-index” Available: http://en.wikipedia.org/wiki/H-index, 2012, viewed on August 9, 9.

S. Wu, J. M. Hofman, W. a. Mason, and D. J. Watts, “Who says what to whom on twitter,” *Proceedings of the 20th international conference on World wide web - WWW ’11*, p. 705–714, 2011.

L. Ya-ting and C. Jing-min, “The social network analysis of political blogs in people: Based on centrality,” *2011 International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pp. 5441–5444, 2011.

J. Yu, “Typology and Influence Analysis of Opinion Leader:A case study on fashion online shopping," *2011 International Conference on E -Business and E -Government (ICEE)*, pp. 1–4, May 2011.

D. Zarrella, “Can Having a Twitter Bio Get You 8 Times as Many Followers?” Available: http://blog.hubspot.com/blog/tabid/6307/bid/4604/Can-Having-a-Twitter-Bio-Get-You-8-Times-as-Many-Followers.aspx, 2009, viewed on August 15, 12

H. Zhou, D. Zeng, and C. Zhang, “Finding Leaders from Opinion Networks,” *IEEE International Conference on Intelligence and Security Informatics, 2009. ISI'09*, pp. 266–268, 2009

APPENDICIES

Appendix A:

The List of Influential Users of the tweets collection, which was a queried on Novermber 5th, 2013, with the words "باسم يوسف":

- **DrBassemYoussef**: A popular Egyptian TV show host; the subject of this collection's topic.
- **Youssefalhosiny**: An Egyptian radio and TV presenter who started out as a political writer.
- **osamagharib1**: He identifies himself as an Egyptian author.
- **amansouraja**: A TV presenter and producer.
- **NaguibSawiris**: A well known businessman, founding member of Al Masreyeen Al Ahrrar political party and owns one of the popular TV channels.
- **Awadalqarni**: A Saudi public figure; an Islamic cleric.
- **FadelSoliman**: An Egyptian muslim apologist, orator, author and film maker, the director of Bridges Foundation.
- **alnagar80**: A known Egyptian activist who may be recognized from appearing on TV and being a former member of parliament.
- **abdrhmanabnody**: An esteemed Egyptian poet. He often makes TV appearances, and is known to voice his views.
- **waleedalfarraj**: A Saudi sports TV presenter.
- **YZaatreh**: A Palestinian author and political analyst.
- **_Andeel_**: An Egyptian cartoonist and script writer.
- **N_AbuBakr**: A young Egyptian writer. Opinionated and often bold.
- **SallamSalah**: A regular user.
- **magdymohamed_**: A regular user.
- **Asmaa2Samir**: A regular user.
- **Sandmonkey**: A regular user.
- **Gemyhood**: A regular user.
- **Salamah**: A regular user.
- **Tahoun71**: A regular user.
- **mo3tzadel**: A regular user.
- **Bassem_Sabry**: A regular user.
- **Gamaleid**: A regular user.
- **Abdelbariatwan**: A regular user.
- **Hmd_Almajed**: A regular user.
- **sofyan_khodary**: A regular user.
- **zaki_safar**: A regular user.
- **hameed_farouq**: A regular user.
- **Aadly_Mansor**: Parody account owned by a regular user.
- **A_Mansorr**: Parody account owned by a regular user.
- **BarackObama_Ar**: Parody account owned by a regular user.

Appendix B:

Table 1: the top 50 users ranked according to their Average Activity Rate

	Screen Name	Average Activity Rate			Screen Name	Average Activity Rate	
1	hesham_m_2011	277.63		26	bibikwt1	120.63	
2	Muhammetusama	231.49		27	aft_7	118.43	
3	magdymohamed_	213.12	Influential	28	himahelaly	110.82	
4	Rawansa3ed	204.32		29	Tahoun71	110.67	Influential
5	Z_o__Z_e	198.18		30	_Mishaall	110.19	
6	nana_25111	194.09		31	h241818	109.9	
7	sara_sara1143	191.33		32	Ala2Atef	108.38	
8	m7md_3abdoo	177.98		33	sasoo_sara1	104.85	
9	JAREDITMISRCOM	160.72		34	A_M_Sabry	103.85	
10	AlesandroAli	159.43		35	abdallahhatem91	102.58	
11	soleman666	157.74		36	youngeagle100	101.74	
12	cawana2013	152.2		37	Muhamed3amr	98.45	
13	chy_jevara	151.42		38	hotm_fa450	98.37	
14	SketrAhmed	135.15		39	Rab4awy	98.36	
15	Oma7R	134.57		40	1_198766	97.49	
16	TarekKamelMoham	133.32		41	a2011abm	97.13	
17	SH_7ezb_Alrayah	132.02		42	fo2fo2_	96.54	
18	miso_ksa	131.52		43	JosphineMamdouh	96.4	
19	quiet_life1417	128.39		44	muhmed002	94.25	
20	asmaa2447	128.37		45	HelpEGY	93.77	
21	aramzy66	128.19		46	omnya821Hawa	92.57	
22	MernaElshap	125.09		47	scarabio7	91.36	
23	Osama_bashaa1	121.87		48	N_AbuBakr	91.29	Influential
24	FinalRule	121.07		49	ThanksShafik	90.8	
25	Ezlam_	120.9		50	control_kw	90.72	

Table 2: the top 50 users ranked according to their Age_Activity Combination scores

	Screen Name	Age_Activity Combination			Screen Name	Age_Activity Combination	
1	hesham_m_2011	0.706		26	JAREDITMISRCOM	0.425	
2	Sandmonkey	0.562	Influential	27	Shrbo	0.423	
3	Gemyhood	0.555	Influential	28	zelaky	0.421	
4	arom4	0.49		29	SketrAhmed	0.421	
5	chy_jevara	0.49		30	Muhammetusama	0.419	
6	nana_25111	0.484		31	FinalRule	0.412	
7	Nawaret	0.478		32	sara_sara1143	0.412	
8	salamah	0.477	Influential	33	soleman666	0.411	
9	ShaimaAboElkhir	0.472		34	ebnmasr	0.408	
10	Tahoun71	0.471	Influential	35	asmaa2447	0.406	
11	magdymohamed_	0.467	Influential	36	Z_o__Z_e	0.405	
12	brhom	0.46		37	Mamdouh_Egypt	0.401	
13	BarackObama_Ar	0.454	Influential	38	Ezlam_	0.397	
14	Bassem_Sabry	0.454	Influential	39	A_M_Sabry	0.393	
15	kazakhelo	0.453		40	Muhamed3amr	0.39	
16	gamaleid	0.45	Influential	41	sotaita7sabo	0.389	
17	hesham9911	0.448		42	Almatrafi	0.387	
18	alnagar80	0.442	Influential	43	ihabtara	0.386	
19	Hazem_Azim	0.441		44	AbdullaAlami	0.385	
20	aramzy66	0.44		45	a2011abm	0.384	
21	MaisAbusalah	0.437		46	amalyou	0.383	
22	almuraisy	0.436		47	EnG_Seif_ElDin	0.383	
23	mariam_yassin	0.431		48	TarekKamelMoham	0.382	
24	RaniaKeiy	0.431		49	elsaudi0	0.381	
25	Rawansa3ed	0.428		50	YasminMahfouz	0.377	

Table 3: the top 50 users ranked according to their Followers Count

	Screen Name	Followers Count			Screen Name	Followers Count	
1	DrBassemYoussef	2162557	Influential	26	Gemyhood	101884	Influential
2	waleedalfarraj	1398721	Influential	27	ma7mod_badr	98743	
3	NaguibSawiris	1076559	Influential	28	Dxbai	91260	
4	Almoslemani	909649		29	MANSOOR_ALJAMRi	87838	
5	Youssefalhosiny	860830	Influential	30	Mounir_Tweets	86438	
6	awadalqarni	749938	Influential	31	Hmd_Almajed	83999	Influential
7	abdrhmanabnody	721614	Influential	32	aboban9	83704	
8	alnagar80	624734	Influential	33	abo3asam	82068	
9	Hazem_Azim	395053		34	assafir	80535	
10	Almatrafi	380347		35	alshaikhmhmd	79198	
11	amansouraja	312548	Influential	36	miso_ksa	78497	
12	gamaleid	299195	Influential	37	YasminMahfouz	66155	
13	abdelbariatwan	287809	Influential	38	N_AbuBakr	64854	Influential
14	rimamaktabi	279677		39	hisham_algakh	62506	
15	AhmedHeImy1811	266190		40	hameedalbloushi	57346	
16	BarackObama_Ar	266009	Influential	41	salamah	52037	Influential
17	khalaf_h	253098		42	Amir3id	49088	
18	badriahalbeshr	244934		43	brhom	48668	
19	FadelSoliman	239275	Influential	44	zelaky	46994	
20	JKhashoggi	232585		45	MustafaSamirE46	46596	
21	YZaatreh	227183	Influential	46	7ely	45220	
22	Sandmonkey	132411	Influential	47	osamagharib1	44907	Influential
23	engyhamdy	117111	Influential	48	OlaOmaar	44488	
24	sayidatynet	102551		49	Bassem_Yossef	42395	
25	Bassem_Sabry	102029	Influential	50	hameed_farouq	40556	Influential

Table 4: the top 50 users ranked according to their TFF Ratio

	Screen Name	TFF ratio			Screen Name	TFF ratio	
1	assafir	26845		26	news_Speed1	510	
2	Almoslemani	18192.98		27	awadalqarni	496.98	Influential
3	NaguibSawiris	8034.02	Influential	28	waleedalfarraj	436.83	Influential
4	MnshorNews	7805		29	khalaf_h	429.71	
5	abdrhmanabnody	7363.41	Influential	30	hassanshahin5	425.05	
6	Aadly_Mansour	3726		31	A_Mansorr	423.31	Influential
7	DrBassemYoussef	3488	Influential	32	sabaia_style	396.97	
8	miso_ksa	3270.71		33	TheAdlyMansour	383.5	
9	AhmedHeImy1811	2559.52		34	Hmd_Almajed	326.84	Influential
10	Mounir_Tweets	2542.29		35	Almatrafi	313.56	
11	alnagar80	1561.84	Influential	36	Shikabala_EGY	252.48	
12	Bassem_Yossef	1413.17		37	osamagharib1	244.06	Influential
13	YZaatreh	1352.28	Influential	38	FadelSoliman	239.28	Influential
14	amansouraja	1255.21	Influential	39	Hazem_Azim	229.68	
15	abdelbariatwan	931.42	Influential	40	sayidatynet	227.89	
16	UBassemoon	906.8		41	control_kw	217.31	
17	Youssefalhosiny	826.93	Influential	42	egynemo	200	
18	badriahalbeshr	790.11		43	meshalfayah	199.82	
19	JKhashoggi	750.27		44	engyhamdy	182.7	Influential
20	3zzMaShkel	684.33		45	MANSOOR_ALJAMRi	182.24	
21	ThanksShafik	669		46	BarackObama_Ar	181.08	Influential
22	Gemyhood	665.91	Influential	47	FilFan	175.73	
23	AhmadMursi	644.38		48	hisham_algakh	163.2	
24	Lotfy_labyb	636.3		49	ma7mod_badr	160.3	
25	rimamaktabi	576.65		50	asadabukhalil	152.8	

Table 5: the top 50 users ranked according to their Collection Tweets Count

	Screen Name	Collection Tweets Count			Screen Name	Collection Tweets Count	
1	m7md_3abdoo	20		26	NaguibSawiris	3	Influential
2	birs3	19		27	YZaatreh	3	Influential
3	HamadaH63721723	8		28	Sandmonkey	3	Influential
4	1_198766	7		29	2t7dawe	3	
5	SketrAhmed	7		30	ahmadayman5	3	
6	N_AbuBakr	6	Influential	31	doaaelsordy	3	
7	DrBassemYoussef	5	Influential	32	abdallah_magdy	3	
8	fadyfikry2	5		33	kaliheragmi	3	
9	assafir	4		34	sara_sara1143	3	
10	badriahalbeshr	4		35	CAP_SHADY	3	
11	Hmd_Almajed	4	Influential	36	kareneenaa1	3	
12	SH_7ezb_Alrayah	4		37	SamrBhettir	3	
13	mo3tzadel	4	Influential	38	awwadwissam	3	
14	rashek_eslami	4		39	amr1771980	3	
15	amro1250	4		40	omnya821Hawa	3	
16	AbdallahBahy	4		41	mikon22	3	
17	zamalkawya57	4		42	Muhammed_Saleim	3	
18	ForConqer	4		43	TarekKamelMoham	2	
19	alokhbaragel	4		44	abdelbariatwan	2	Influential
20	algohiny	4		45	JKhashoggi	2	
21	AhmedFayez34	4		46	AhmadMursi	2	
22	yahya_zekaa	4		47	Lotfy_labyb	2	
23	mohammed_hagag_	4		48	Shikabala_EGY	2	
24	sh1614	4		49	osamagharib1	2	Influential
25	Safaa_44	4		50	meshalfayah	2	

Table 6: the top 50 users ranked according to their Average Retweet Counts

	Screen Name	Average Retweets Count			Screen Name	Average Retweets Count	
1	Youssefalhosiny	1286	Influential	26	abdrhmanabnody	160	Influential
2	osamagharib1	608.5	Influential	27	ahmedshabana94	148	
3	N_AbuBakr	533.3	Influential	28	waleedalfarraj	145	Influential
4	amansouraja	459	Influential	29	Eslam_Luca	139.5	
5	DrBassemYoussef	443.2	Influential	30	2t7dawe	138.7	
6	TheAdlyMansour	427		31	drGABER_NASSAR	137	
7	SallamSalah	388	Influential	32	Adhamabdelshafy	135	
8	Amir3id	386		33	YZaatreh	132	Influential
9	RebelNabil	359		34	safwanmohamed	132	
10	iromyys	318		35	rabawy7	130	
11	AhmedHeImy1811	296		36	khalaf_h	124	
12	alnagar80	291	Influential	37	ahmedmontie96	119.5	
13	Almoslemani	266		38	JKhashoggi	117.5	
14	JosphineMamdouh	248		39	MostafaManno	117	
15	ma7mod_badr	246		40	AhmadMursi	108	
16	MontaserMarai	242		41	MhamedKrichen	108	
17	mo3tzadel	233.8	Influential	42	dianamoukalled	107	
18	Asmaa2Samir	231	Influential	43	NaguibSawiris	105	Influential
19	youssefamr1996	226		44	Erhabawi	101	
20	Mostafa_T_Awny	224		45	sofyan_khodary	96	Influential
21	awadalqarni	200	Influential	46	Mina_Sha7toty	93.5	
22	Lotfy_labyb	198.5		47	abdelbariatwan	91.5	Influential
23	_Andeel_	190	Influential	48	Rawansa3ed	86	
24	alobisan	175		49	Gemyhood	83	Influential
25	Mounir_Tweets	164		50	7ely	82	

Table 7: the top 50 ranked according to their Average Tweets' Age (in minutes)

	Screen Name	Average Tweets Age			Screen Name	Average Tweets Age	
1	freenanno	280463		26	A_M_Sabry	13862	
2	A_3adl	185786		27	7assan_z	13725	
3	ameena_alkuwari	89707		28	7ely	13126	
4	mohamed_alaa14	58365		29	AMagdiZ	13078	
5	DrBassemYoussef	34827	Influential	30	scarabio7	12745	
6	MrSuspended	20299		31	Sheexo	12626	
7	FadelSoliman	18719	Influential	32	zyazigi	11879	
8	spoony___	15824		33	M7Slama	11752	
9	DaliaFaisalL	15644		34	abdallahhatem91	11306	
10	fo2fo2_	15509		35	mahysafwat	11253	
11	she3aa14	15437		36	asoomcr7	10910	
12	AbdallahBahy	15422		37	EssamMuhammadd	10561	
13	lithymohamed	15419		38	Adhamabdelshafy	10398	
14	Engy_Ahmed98	15393		39	SallamSalah	10349	Influential
15	MazenAlosali	15391		40	AhmadMursi	10298	
16	NadineYosry	15369		41	osamagharib1	10295	Influential
17	ahmedsamehmuhmd	15369		42	k0oz	9953	
18	sofyan_khodary	15362	Influential	43	SarahElGandour	9747	
19	Muhamed3amr	15340		44	JAWAHER_ALSAIF	9464	
20	zaki_safar	15137	Influential	45	ShaimaAboElkhir	8853	
21	Z_o__Z_e	15089		46	i_aryam	8814	
22	KaremM7md	15012		47	5764464	8722	
23	selvianaguib	14408		48	2t7dawe	8720	
24	wooda2000	14019		49	N_AbuBakr	8716	Influential
25	iMayooda	14010		50	ASHRAFel_MAHDY	8200	

Table 8: the top 50 users ranked according to their Average Retweet Frequency

	Screen Name	Average Retweet Frequency			Screen Name	Average Retweet Frequency	
1	Ezlam_	0.3		26	YZaatreh	0.044	Influential
2	Youssefalhosiny	0.239	Influential	27	mo3tzadel	0.044	Influential
3	Aadly_Mansor	0.15	Influential	28	Asmaa2Samir	0.043	Influential
4	martinamedhat16	0.133		29	abdrhmanabnody	0.043	Influential
5	A_Mansorr	0.11	Influential	30	youssefamr1996	0.042	
6	amansouraja	0.105	Influential	31	Mostafa_T_Awny	0.042	
7	Be3are	0.095		32	alobisan	0.04	
8	TheAdlyMansour	0.08		33	dianamoukalled	0.038	
9	Mohamed_tottie	0.079		34	Lotfy_labyb	0.038	
10	ArchLucy	0.077		35	khalaf_h	0.038	
11	Amir3id	0.072		36	SallamSalah	0.037	Influential
12	RebelNabil	0.067		37	LailaAbdElRaof7	0.036	
13	N_AbuBakr	0.061	Influential	38	messelhi	0.034	
14	iromyys	0.059		39	waleedalfarraj	0.033	Influential
15	_Andeel_	0.059	Influential	40	zelaky	0.032	
16	osamagharib1	0.059	Influential	41	Mounir_Tweets	0.032	
17	AhmedHeImy1811	0.058		42	Moliimoll	0.031	
18	alnagar80	0.055	Influential	43	gamaleid	0.028	Influential
19	awadalqarni	0.055	Influential	44	ahmedshabana94	0.027	
20	Almoslemani	0.051		45	alqaheraalyoom	0.027	
21	salamah	0.048	Influential	46	alshaikhmhmd	0.027	
22	ma7mod_badr	0.047		47	JKhashoggi	0.027	
23	asadabukhalil	0.046		48	hameed_farouq	0.027	Influential
24	JosphineMamdouh	0.046		49	Eslam_Luca	0.026	
25	MontaserMarai	0.045		50	drGABER_NASSAR	0.026	

Table 9: the top 50 users ranked according to score 1

	Screen Name	score 1			Screen Name	score 1	
1	Youssefalhosiny	0.525	Influential	26	Asmaa2Samir	0.099	Influential
2	DrBassemYoussef	0.322	Influential	27	YZaatreh	0.094	Influential
3	Ezlam_	0.285		28	youssefamr1996	0.094	
4	freenanno	0.25		29	MontaserMarai	0.093	
5	amansouraja	0.206	Influential	30	mo3tzadel	0.092	Influential
6	waleedalfarraj	0.193	Influential	31	ameena_alkuwari	0.091	
7	N_AbuBakr	0.189	Influential	32	hesham_m_2011	0.089	
8	Almoslemani	0.179		33	Mohamed_tottie	0.089	
9	osamagharib1	0.178	Influential	34	_Andeel_	0.088	Influential
10	A_3adl	0.172		35	Be3are	0.088	
11	alnagar80	0.164	Influential	36	Mostafa_T_Awny	0.087	
12	awadalqarni	0.161	Influential	37	Rawansa3ed	0.084	
13	TheAdlyMansour	0.152		38	ArchLucy	0.083	
14	Amir3id	0.143		39	khalaf_h	0.082	
15	NaguibSawiris	0.14	Influential	40	Mounir_Tweets	0.077	
16	Aadly_Mansor	0.138	Influential	41	Lotfy_labyb	0.077	
17	abdrhmanabnody	0.134	Influential	42	Hazem_Azim	0.076	
18	RebelNabil	0.134		43	alobisan	0.075	
19	AhmedHeImy1811	0.132		44	Eslam_Luca	0.075	
20	iromyys	0.13		45	JKhashoggi	0.073	
21	martinamedhat16	0.129		46	Z_o__Z_e	0.07	
22	JosphineMamdouh	0.118		47	abdelbariatwan	0.069	Influential
23	SallamSalah	0.117	Influential	48	Almatrafi	0.067	
24	A_Mansorr	0.103	Influential	49	abo3asam	0.067	
25	ma7mod_badr	0.102		50	gamaleid	0.066	Influential

Table 10: the top 50 users ranked according to score 2

	Screen Name	score 2			Screen Name	score 2	
1	Youssefalhosiny	0.434	Influential	26	ma7mod_badr	0.084	
2	DrBassemYoussef	0.415	Influential	27	Asmaa2Samir	0.084	Influential
3	freenanno	0.334		28	Almatrafi	0.08	
4	A_3adl	0.23		29	youssefamr1996	0.079	
5	waleedalfarraj	0.22	Influential	30	YZaatreh	0.076	Influential
6	N_AbuBakr	0.184	Influential	31	mohamed_alaa14	0.076	
7	Almoslemani	0.182		32	FadelSoliman	0.076	Influential
8	osamagharib1	0.172	Influential	33	magdymohamed_	0.074	Influential
9	NaguibSawiris	0.163	Influential	34	mo3tzadel	0.074	Influential
10	amansouraja	0.158	Influential	35	MontaserMarai	0.073	
11	alnagar80	0.157	Influential	36	nana_25111	0.071	
12	awadalqarni	0.153	Influential	37	Eslam_Luca	0.07	
13	abdrhmanabnody	0.132	Influential	38	Mostafa_T_Awny	0.069	
14	ameena_alkuwari	0.121		39	Muhamed3amr	0.068	
15	SallamSalah	0.114	Influential	40	khalaf_h	0.068	
16	TheAdlyMansour	0.113		41	Mounir_Tweets	0.068	
17	AhmedHeImy1811	0.112		42	JKhashoggi	0.067	
18	Amir3id	0.11		43	abdelbariatwan	0.067	
19	iromyys	0.108		44	sofyan_khodary	0.064	Influential
20	JosphineMamdouh	0.107		45	sara_sara1143	0.063	
21	hesham_m_2011	0.105		46	abo3asam	0.061	
22	RebelNabil	0.104		47	Lotfy_labyb	0.06	
23	Rawansa3ed	0.094		48	Ala2Atef	0.06	
24	Hazem_Azim	0.092		49	Adhamabdelshafy	0.059	
25	Z_o__Z_e	0.09		50	A_M_Sabry	0.059	

Table 11: the top 50 users ranked according to score 3

	Screen Name	score 3			Screen Name	score 3	
1	Youssefalhosiny	0.69	Influential	26	mo3tzadel	0.111	Influential
2	DrBassemYoussef	0.385	Influential	27	YZaatreh	0.11	Influential
3	Ezlam_	0.335		28	Asmaa2Samir	0.107	Influential
4	amansouraja	0.269	Influential	29	Be3are	0.107	
5	waleedalfarraj	0.242	Influential	30	youssefamr1996	0.106	
6	Almoslemani	0.233		31	Mohamed_tottie	0.105	
7	osamagharib1	0.223	Influential	32	Mostafa_T_Awny	0.104	
8	alnagar80	0.21	Influential	33	khalaf_h	0.103	
9	N_AbuBakr	0.209	Influential	34	Lotfy_labyb	0.095	
10	awadalqarni	0.201	Influential	35	alobisan	0.088	
11	TheAdlyMansour	0.196		36	JKhashoggi	0.087	
12	Aadly_Mansor	0.183	Influential	37	Mounir_Tweets	0.087	
13	Amir3id	0.182		38	ArchLucy	0.086	
14	NaguibSawiris	0.18	Influential	39	abdelbariatwan	0.083	Influential
15	abdrhmanabnody	0.174	Influential	40	gamaleid	0.077	Influential
16	AhmedHeImy1811	0.17		41	dianamoukalled	0.071	
17	RebelNabil	0.164		42	ahmedshabana94	0.069	
18	martinamedhat16	0.152		43	Almatrafi	0.069	
19	iromyys	0.146		44	Hazem_Azim	0.067	
20	SallamSalah	0.139	Influential	45	Eslam_Luca	0.065	
21	A_Mansorr	0.136	Influential	46	drGABER_NASSAR	0.064	
22	ma7mod_badr	0.126		47	safwanmohamed	0.061	
23	JosphineMamdouh	0.116		48	rabawy7	0.059	
24	_Andeel_	0.114	Influential	49	salamah	0.059	Influential
25	MontaserMarai	0.111		50	asadabukhalil	0.057	

Table 12: the top 50 users ranked according to score 4

	Screen Name	score 4			Screen Name	score 4	
1	Youssefalhosiny	0.637	Influential	26	youssefamr1996	0.088	
2	DrBassemYoussef	0.557	Influential	27	abdelbariatwan	0.086	Influential
3	waleedalfarraj	0.307	Influential	28	JKhashoggi	0.086	
4	Almoslemani	0.264		29	Mostafa_T_Awny	0.086	
5	osamagharib1	0.236	Influential	30	Hazem_Azim	0.086	
6	NaguibSawiris	0.234	Influential	31	Lotfy_labyb	0.079	
7	amansouraja	0.228	Influential	32	Mounir_Tweets	0.077	
8	alnagar80	0.222	Influential	33	_Andeel_	0.072	Influential
9	N_AbuBakr	0.211	Influential	34	FadelSoliman	0.071	Influential
10	awadalqarni	0.21	Influential	35	gamaleid	0.069	Influential
11	abdrhmanabnody	0.19	Influential	36	alobisan	0.066	
12	TheAdlyMansour	0.16		37	BarackObama_Ar	0.064	Influential
13	AhmedHeImy1811	0.159		38	rimamaktabi	0.059	
14	Amir3id	0.153		39	ahmedshabana94	0.058	
15	SallamSalah	0.147	Influential	40	Eslam_Luca	0.054	
16	RebelNabil	0.135		41	2t7dawe	0.053	
17	iromyys	0.12		42	drGABER_NASSAR	0.053	
18	ma7mod_badr	0.11		43	Adhamabdelshafy	0.052	
19	JosphineMamdouh	0.097		44	badriahalbeshr	0.051	
20	mo3tzadel	0.093	Influential	45	safwanmohamed	0.05	
21	khalaf_h	0.092		46	Gemyhood	0.049	Influential
22	MontaserMarai	0.091		47	rabawy7	0.049	
23	YZaatreh	0.09	Influential	48	ahmedmontie96	0.048	
24	Asmaa2Samir	0.089	Influential	49	MostafaManno	0.044	
25	Almatrafi	0.088		50	MhamedKrichen	0.043	

Table 13: the top 50 users ranked according to the average Perplexity value of their tweets

	Screen Name	Average Perplexity			Screen Name	Average Perplexity	
1	shehabkhaledd	9.377		26	MOHAMED_SHARAF_	28.117	
2	Mo7amedSala7_	9.377		27	MasrElyoom	28.154	
3	ahmedmelbanna	9.377		28	omar_k3	31.971	
4	SuzanYoussef	9.377		29	Almoslemani	32.630	
5	Bassem_Sabry	9.377	Influential	30	_Andeel_	32.802	Influential
6	israabasha1	9.377		31	o_m77	33.349	
7	Meedoo_YJ	9.377		32	waleedalfarraj	36.150	Influential
8	thanks_me	9.377		33	Mayarelfadaly	40.035	
9	SHeKooSNiPeR	9.377		34	rahman2267	40.666	
10	Nour_salah0	9.377		35	JosphineMamdouh	42.883	
11	samandaImaher	9.377		36	Rawansa3ed	42.996	
12	pinky_jojo1	9.377		37	RanaMohamedd	42.996	
13	iromyys	9.377		38	Eslam_Luca	43.509	
14	A_3adl	9.377		39	ameena_alkuwari	43.862	
15	RaaefGad	9.377		40	amrmohkhalifa	45.443	
16	HALN3IMI	9.377		41	aws_89	45.674	
17	mahysafwat	9.377		42	TheAdlyMansour	48.976	
18	LaamelIliace	9.377		43	Ezz_1907	53.422	
19	Montherabdulah	10.623		44	MostaFaChika1	55.017	
20	O_00_O_	18.706		45	Egypt_SS	56.136	
21	alobisan	22.586		46	hatemamen	56.302	
22	khalaf_h	24.663		47	ShroukRashwan	56.443	
23	Bassem_Yossef	27.154		48	badriahalbeshr	57.865	
24	Sheexo	27.352		49	mtito9245	57.865	
25	RanaBadr46	27.352		50	tay_Koo	58.512	

Table 14: the top 50 users ranked according to the minimum Perplexity value of their tweets

	Screen Name	Minimum Perplexity			Screen Name	Minimum Perplexity	
1	shehabkhaledd	9.377		26	HamadaH63721723	24.212	
2	Mo7amedSala7_	9.377		27	khalaf_h	24.663	
3	ahmedmelbanna	9.377		28	Bassem_Yossef	27.154	
4	SuzanYoussef	9.377		29	Mina_Sha7toty	27.266	
5	Bassem_Sabry	9.377	Influential	30	Sheexo	27.352	
6	israabasha1	9.377		31	RanaBadr46	27.352	
7	Meedoo_YJ	9.377		32	MOHAMED_SHARAF_	28.117	
8	thanks_me	9.377		33	MasrElyoom	28.154	
9	SHeKooSNiPeR	9.377		34	Blacklist25Jan	28.154	
10	Nour_salah0	9.377		35	Bassemlovers	28.154	
11	samandaImaher	9.377		36	N_AbuBakr	30.050	Influential
12	pinky_jojo1	9.377		37	omar_k3	31.971	
13	iromyys	9.377		38	YZaatreh	31.971	Influential
14	A_3adl	9.377		39	Almoslemani	32.630	
15	RaaefGad	9.377		40	_Andeel_	32.802	Influential
16	HALN3IMI	9.377		41	o_m77	33.349	
17	mahysafwat	9.377		42	waleedalfarraj	36.150	Influential
18	LaamelIliace	9.377		43	Eslam_Luca	37.758	
19	AhmedFayez34	9.377		44	Mayarelfadaly	40.035	
20	MriiemWael	9.377		45	rahman2267	40.666	
21	Montherabdulah	10.623		46	JosphineMamdouh	42.883	
22	O_00_O_	18.706		47	Rawansa3ed	42.996	
23	DrBassemYoussef	20.201	Influential	48	RanaMohamedd	42.996	
24	alobisan	22.586		49	ameena_alkuwari	43.862	
25	mo3tzadel	24.212	Influential	50	osamagharib1	44.949	Influential