Semantic Analysis of Micro-blogs

By

Umar Hayat Khan Niazi

2007-NUST-MS-PhD IT-33

Supervisor

Dr. Khalid Latif

A thesis submitted in partial fulfillment of the requirements for the degree of

Masters of Science in Information Technology (MSIT)

in

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan

(March 2011)

APPROVAL

It is certified that the contents and form of thesis entitled **"Semantic Analysis of Micro-blogs"** submitted by **Umar Hayat Khan Niazi** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Khalid Latif**

Signature: _____

Date: _____

Committee Member 1: **Dr. Sharifullah Khan**

Signature: _____

Date: _____

Committee Member 2: **Dr. Zia ul Qayyum**

Signature: _____

Date: _____

Committee Member 3: **Mr. Osama Hashmi**

Signature: _____

Date: _____

To my Parents, Wife, Children and Teachers for their great love and support throughout this research and my career

CERTIFICATE OF ORIGINALITY

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by any other person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Umar Hayat Khan Niazi**

Signature: _____

ACKNOWLEDGEMENTS

First and foremost, I am immensely thankful to Almighty Allah for letting me pursue and fulfill my dreams. Nothing could have been possible without His blessings.

It is with pleasure that I express my loving gratitude for my thesis supervisor Dr. Khalid Latif, for his kind supervision, guidance, and advice from the very early stages of this research. Much of what lies in the following pages can be credited to his kind supervision. His truly scientist hunch and passions in research exceptionally inspire and enrich my growth as a student and a researcher. Above all and the most needed, he provided me persistent encouragement and support in numerous ways. I am indebted to him more than he knows.

I am also extremely thankful to my committee members Dr. Sharifullah Khan, Dr. Zia ul Qayyum and Mr. Osama Hashmi for their support, valuable suggestions, and positive criticism. My parents and my wife deserve special mention for all their love, encouragement, inseparable support and prayers.

Finally, I would like to record my thanks for my friends, and colleagues especially at DELSA lab for their enormous encouragement and help, which made my life more pleasant and easier during research work. Let me grab this opportunity to thank everybody who was important to the successful realization of thesis, as well as expressing my apology that I could not mention personally one by one.

Umar Hayat Khan Niazi

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

API:	Application Programmer Interface
AI:	Artificial Intelligence
CAPTCHA:	Completely Automated Public Turing test to tell Computer and Humans Apart
IM:	Instant Message
IP:	Internet Protocol
JSON:	JavaScript Object Notation
NE:	Named Entity
NLP:	Natural Language Processing
PHP:	Hypertext Preprocessor
RDF:	Resource Description Framework
RT:	Re-Tweet
SEO:	Search Engine Optimization
SMS:	Short Message Service
SVM:	Support Vector Machine
UBE:	Unsolicited Bulk Email
URL:	Uniform Resource Locator
XML:	Extensible Markup Language

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Micro-blogging platforms have proven their importance as vital communication channels over the internet. Individuals use micro-blogging platforms to keep in touch with friends and families whereas corporate users make use of it to introduce new products and services to their clients. Spammers also cash in on the global reach of micro-blogs to spread irrelevant, immaterial and offensive stuff like viruses, porn etc. Spammers are wasting resources, valued user time and annoying valid users by polluting these platforms with their orthogonal messages. Identifying an irrelevant message on such platforms is a challenging task. A user sending legitimate messages most of the times and infrequently sending junk replies cannot be declared as a spammer. Similarly, public messages, such as advertisements, can be considered irrelevant by one reader but relevant by another due to their diverse personal interests. These messages contain named entities, URLs, events, facts and figures. These named entities have different relationships among them. With the current, state of the art semantic information extraction and analysis techniques it has become possible to dig out these named entities and their relationships with each other. In this research we have implemented an algorithm to detect the irrelevant messages on one of the famous micro-blogging platforms known as Twitter. Our algorithm utilizes the semantic information extraction and analysis techniques to compute relevance among different parts of the messages and compares it with a user set threshold. The messages with higher similarity among their components are most likely the relevant message and vice versa. We have validated our algorithm to detect irrelevant messages from a dataset collected from Twitter. Our algorithm has successfully achieved a precision of up to 97% with equally good values for recall and F-Measure up to 100% and 97% respectively.

CHAPTER 1

INTRODUCTION

Micro-blogging allows users to post brief text updates, links or media such as photos or audio clips. It allows rapid content publishing and information distribution. At 1,382%, Twitter [1] is fastest growing social networking & micro-blogging platform with Zimbio[1] and Facebook[2] subsequently at 240% and 228% [2].

Twitter is one of the popular micro-blogging platform. It allows users to keep in touch with other users through exchange of quick, frequent messages called tweets. Total number of characters allowed in a tweet is 140. Tweets can be sent directly from Twitter website or a range of other clients and services such as Short Messaging Service (SMS) devices like mobile phones and Instant Messaging (IM) tools like Google Talk[3]. Twitter users track each other to build their social networks. Users can re-distribute contents from their blogs to the masses in real time. Journalists are pitching rapidly emerging stories on Twitter whereas corporations are utilizing it for brand awareness and getting direct response marketing. Dell[4], for instance has its own Twitter account that has generated $3 million sales since 2007[3]. Similarly Moonfruit [5] campaign on Twitter has brought in more than $30,000 in a single month for a relatively small and less famous company.

According to a study performed at Penn State University, 20% of the tweets are either an inquiry or information about a specific product or service [4]. Tweets contain important information such as Named Entities (NEs). The NEs are the items of highest interest on web and in 2004, all top 10 searches on Yahoo were named entities [7]. NEs have relationships among each other. These relationships can be extracted and thus utilized in finding the

[1] http://www.zimbio.com/
[2] http://www.facebook.com/
[3] http://www.google.com/talk/
[4] http://www.dell.com
[5] http://www.moonfruit.com/

semantic relatedness between different elements of tweets [8] and to filter out irrelevant tweets. There are many traditional information extraction techniques used to extract named entities, the context and the relationships among them. Semantic web initiative by W3C[6] is an extension of the current web to bring web information in a canonical structure. The ultimate focus is enabling computer and people to work in co-operation [9]. The rich metadata (semantic annotations) of tweets make them a part of information-sharing ecosystem.

1.1 Motivation

Many corporations are executing their marketing campaigns on Twitter. These campaigns give out money and other highly attractive prizes to users who participate in these campaigns. The users add advertising company's hashtag in their tweets to become a participant of these competitions. Moonfruit is a similar campaign in which Sitemaker Software Limited [7] ("Sitemaker") announced a competition on the eve of its 10th birthday anniversary. This competition lasted for 7 days and they gave away 10 Macbook Pro. The participating candidates were required to include #Moonfruit hashtag in their tweets. The users were free to re-tweet "*Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC#Moonfruit"* [5] or create their own tweets with #Moonfruit. This competition attracted a large number of Twitter users. Statistics of this campaign claim an increase of 600% in traffic and 350% increase in signup and trials on Moonfruit.com [6]. Hashtag #Moonfruit remained number one topic on Twitter for two days. Although it was not necessary to follow company's Twitter user yet the number of followers boosted to 44,113 in only 7 days which shows an increase of approximately 100 times compared to 444 followers. The #Moonfruit hashtag later got hijacked by spammers and they started spreading their own

[6] http://www.w3.org/2001/sw/
[7] http://www.sitemakerlive.com

information with this hashtag in their tweets. They started promoting 'get rich quick' and similar schemes. The hashtag was used for some charitable causes as well. Moonfruit campaign was reduced from 10 days to 7 days in an attempt to diffuse negative sentiment and reduce Twitter feed 'pollution' [6]. Most of the users added hashtag regardless of the relevance to their tweets. Others just re-tweeted and added irrelevant URLs and multiple hashtags. This resulted in no contribution towards the company goals. However a storm of tweets started and caused problems that were analogous to spamming.

1.2 Problem Definition

Spammers are using exploits available on Twitter for spreading spam messages. Most prominent exploits include URL Shortening, hashtags on trending topics, user accounts hijacking and tweet-jacking.

1.2.1 URL Shortening

Users use URL shortening services to shorten a URL. This save spaces and provides more characters for broadcasting the original message. These short URLs are obfuscated and a user cannot know where the URL will take him unless he clicks it. Spammers are taking advantage of this exploit and spreading URLs of their choice and making the Twitter users click on these URLs. In this way, they are distracting the Twitter users. For example, consider a URL *"http://bit.ly/n0iuI"*, one cannot foresee about the URL it is linked to. Users have no idea unless they click it.

1.2.2 Hashtags on Trend Topics

Twitter community adopts a way for embedding additional context and metadata to the Twitter messages. This new way is called hashtag. Hashtag can be created by simply

prefixing a word with a hash symbol: #hashtag. The hashtags are just like the social tags on other social networking and blogging sites. However these hashtags are only added inline to the post. Twitter provides analytic reports and indexing features on these hashtags to allow users to track what's happening now. Based on these analytic reports and statistics, Twitter displays hot topics on a very prominent place on its website. These hot topics on Twitter are called trending topics. One can include trending topic's hashtag in tweets to make Twitter to include his tweet in trending topic's tweet stream. Whenever a search for this trending topic hashtag is made, these tweets are shown in the search results. Spammers are taking advantage of this exploit and they include the hashtag into their spam tweets. In this way they are able to increase the visibility of their tweets because these tweets show up in most popular searches. e.g: "Learn something about yourself - Take the FREE psychological test on this great #Moonfruit site – *http://bit.ly/n0iuI"*.

1.2.3 Twitter Account Hijacking

Spammers hijack Twitter user accounts having large number of followers and start sending messages to all the followers of that account. The followers of original account now started to consider the spammers account as original user account and start clicking the URLs sent by the spammers in his tweets.

1.2.4 Tweet-Jacking

Twitter provides a facility to send direct messages to others users by simply including @username in the tweet. These tweets are called direct messages. These tweets are shown in user's dashboard. By combining the URL shortening exploit with this, spammers replace the original URL present in the tweet with their own URL and broadcast the tweet to the community by re-tweeting it. The users might click on this flimsy URL which leads to a

malware or site of spammer's choice. This method of spreading spam is also called convo-spamming.

Thus Twitter users are sending tweets very frequently but most of these tweets are not contributing anything positive for the community. Most of the time instead of creating new tweets many users are re-tweeting others tweets and most of these tweets does not have any relevance among their components. Some users are sharing annoying and disturbing links such as pornographic material. Thus this is causing spam problem on Twitter in the form of irrelevant tweets.

Marketing campaigns are being run on Twitter everyday and these all are creating issues like #Moonfruit campaign did. This and similar kind of other spamming activities are causing wastage of time and important resources. Moreover, it is causing many other problems including:

- Legitimate tweets cannot stay on top of the search for a long time
- Legitimate tweets retrieval becomes very difficult
- Potential tweet retrieval becomes impossible
- Server processing is wasted and misused
- Link Spamming distracts Twitter users

Thus the objective of this research is to discuss the design and development of the algorithm to discover spam out of short text messages on micro-blogging platforms specially Twitter. Major challenges being faced include:

1. Public messages, such as advertisements, can be considered as spam by one user but legitimate by other user due to their diverse personal interests.

2. A user who is sending legitimate messages most of the time and infrequently sending junk replies cannot be adjudged as spammer.
3. How to measure relevance?
4. Which factors to be considered for measuring relevance?
5. Can semantic analysis of the tweet be helpful?

1.3 Proposed Approach

The goal of this research is to design an algorithm for detection of spam tweets on Twitter. It is suggested that the tweets with higher dissimilarity among the components (e.g. hashtags, URLs, and named entity mentions) are more likely spam. Thus the algorithm needs to be based on analysis of tweet contents. A tweet consists of maximum of 140 characters and it could contain following optional elements:

1. Text
2. One or more hashtags
3. One or more URLs
4. One or more User mentions and RT

Twitter users start a topic for discussion and Twitter allows tracking of all tweets on the topic using hashtags. Users reply each other using @username convention and republish the tweets of other users using RT @username in start of their tweets. URLs are included in the tweets for spreading contents from other sites. After doing detailed analysis of tweet contents it is suggested that algorithm should take following factors into account:

1. Semantic relatedness among tweet text, hashtag details and URL title
2. A dynamic list of spam words added by the users

3. Ranking of the Twitter users from Twitter Grader [10]

Each of these factors carries weight and must be considered while making a decision about the tweet status. These factors are described below:

1.3.1 Semantic Relatedness

Calculation of semantic relatedness between:

- 🕮 Tweet text and hashtag details
- 🕮 Tweet text and URL title
- 🕮 URL title and hashtag details

Hashtag details, URL title and Tweet text are in the form of short text segments. Measuring semantic relatedness between short text segments can be best achieved by using the query expansion. Query expansion using Wikipedia articles has been utilized by E. Gabrilovich, S. Markovitch [11]. Similarly Wen-tau & Meek [12] took advantage of the query expansion using search engine results to find the semantic relatedness between the short text segments. These short text segments are expanded with the help of search engine results and then measure the semantic relatedness between them in the expanded universe. This factor is given some weight in order to make a decision about the tweet status.

1.3.2 List of Spam Words

System will maintain a list of commonly used words in spamming activities in a day. The tweets will be tested for these words. This saves us against the strange use of abbreviation by the spammers [13]. This list is open for user submissions and thus it will keep on increasing. If a tweet contains words from this list then this factor contributes accordingly.

1.3.3 Ranking of Users from Twitter Grader

Twitter grader measures power, reach and authority of the Twitter users. It uses following parameters to measure the rank of a Twitter user

- Number of Followers
- Power of Followers
- Tweets frequency
- How recent the tweets are?
- Follower/Following Ratio
- Re-tweets of user's tweets

This ranking is readily accessible via an Application Programmers Interface API. If a user has a high ranking on Twitter grader then score from this factor will not affect much in decision making about the tweet status and vice versa.

The individual scores from the above-described factors are summed up and then compared with a threshold value set by the user to make the final decision about the tweet status. If the total score is less than or equal to the threshold value then the tweet is considered as an irrelevant or spam tweet otherwise the tweet is declared as legitimate tweet for that specific user.

1.4 Thesis Outline

The rest of the thesis document is structured as follows: **Chapter 2** sets the stage by providing the background knowledge about the subject of semantic analysis, spam detection and similarity measures. It also contains literature survey about spam detection and describes different techniques utilized for spam detection in other fields such as email and blogging.

Chapter 3 describes the methodology adopted to detect spam in micro-blogging in detail. A comprehensive overview of the system implementation and results achieved are presented in **Chapter 4**. **Chapter 5** presents the conclusion of the study and does provide an outlook to future research work.

CHAPTER 2

LITERATURE REVIEW

Communication through email, blogs, instant messages, social networks and discussion forums is one of the vital advancements of the web era. Individuals use these channels to keep in touch with each other. Corporate sector users are taking advantage of these channels to introduce their products to a large number of audiences worldwide. Spammers cash in on the global reach of these communication channels to spread irrelevant, immaterial and offensive material like viruses, porn etc. This causes wastage of valued user time, storage and bandwidth. Researchers have made sound efforts to deal with the spam problem in past.

Section 2.1 presents an over view of the communication over the internet. Section 2.2 discusses spam email communication and the methods being used to fight against it. This section tells about spamming on blogs, forums and social networking sites. In this section I will also describe the available ways to eliminate or at least minimize the spamming on this kind of communication channel. Section 2.3 is dedicated to spamming on micro-blogging websites and we will talk about Twitter and spamming on Twitter specifically. In Section 2.4 we will look critically at some research works already done in this field.

2.1 Communication over Internet

Internet has grown rapidly since its growth and transformed the world into a global village by providing new methods of communication. It has emerged as a very powerful platform for doing personal, business and organizational communications over the past two decades. The cost of communication has been lowered. Internet has emerged as a universal resource of information for everyone. Number of internet users continues to grow briskly and according to Bill Clinton, “When I took office, only high energy physicists had ever heard of what is called the World Wide Web. Now even my cat has it's own page.” [14]. Even for personal use

or organizational use, it has become a part of our lives - from communicating with friends to doing a business deal, to applying for a new job or for almost any other transaction we do online in our daily life.

Internet provides public as well as private channels for communication. These channels are used for personal as well as organizational communication purposes. People are using these methods for job applications, business deals, introducing new services, approaching potential customer and following them up for making just another e-commerce transaction. Public ways to communicate include blogging, discussion forums, social networking and community websites. People are discussing ideas, products, services and personalities of their interest on public communication channels round the globe. The public communication channels also provide the facility to send private messages to others.

2.1.1 Spamming in Communication Channels

With all the advancement in communication methods, people exploit these communication channels by spreading un-authorized and unwanted messages. These messages are in bulk and contain URLs to virus, pornographic website, , health care products and other offensive materials. Exploiters continuously send irrelevant and unauthorized messages on both private as well as public channels. This causes users to put extra efforts to segregate the legitimate messages from these extraneous messages. This also wastes server storage and valuable bandwidth. These irrelevant messages are called SPAM. Technically a spam is an electronic message in which recipient's personal identity and context are irrelevant and the message is equally applicable to many recipients and has been sent to them without their consent [15].

2.1.2 Spamming in Email

Today, electronic mail (e-mail) is the most widely mean for communication used by individuals as well as professionals for personal and business/official communications. Different people use email for communicating with friends, applying for jobs, tasks management, business deals and almost in every facet of their lives including the signing up on social networking sites. People use their email addresses to create individual identities on these sites because email addresses are unique. At the same time email spamming has made email systems a headache. Email spamming is also called Unsolicited Bulk Email (UBE) refers to sending emails to hundreds or thousands of users simultaneously for different purposes. It is characterized by abusers repeatedly sending an email message to a particular address at a specific victim site. In many instances, the messages will be large and constructed from meaningless data in an effort to consume additional system and network resources. Multiple accounts at the target site may be abused, increasing the denial of service impact. It has caused severe problems in terms of lost productivity, wastage of bandwidth, administration of network systems and invasion of privacy of users [16]. When large amounts of email are directed to or through a single site, the site may suffer a denial of service through loss of network connectivity, system crashes, or failure of a service because of:

- Overloading network connections
- Using all available system resources
- Filling the disk as a result of multiple postings and resulting syslog entries

The problem of spam detection has been addressed by many data mining researchers. They have treated the spam detection a static text classification problem, but email spamming is

almost impossible to prevent because a user with a valid email address can spam any other valid email address, newsgroup, or bulletin-board service. The popularity and importance of the problem can be well understood from its inclusion in the Data Mining Cup Contest [17].

Filtering a spam message is a typical classification problem. Most spam filtering techniques use text categorization methods. Many researchers have tried to invade this problem and suggested collaborative and content-analysis techniques for spam filtering. Two methods of machine classification are distinguished. One is rule based and the other one is done with the help of machine learning techniques. In rule base method, rules are defined manually when all data classes are static and they can be separated easily based on the features, whereas in machine learning method is applied when the features for differentiating a spam and legitimate messages are not distinct enough.

Hidalgo [18] has discussed text categorization methods for UBE filtering. They utilized and evaluated a number of machine learning methods including C4.5, Naïve Bayes, PART, Support Vector Machines (SVM) and Rochhio. Drucker et.al [19] used support vector machines for spam categorization. They devised a learning machine and either color-coded the spam messages for user or presented them in the order of degree of confidence. They left the final decision to the user to mark these color-coded or low ranked messages as spam or legitimate. Artificial Intelligence (AI) is also being used to fight against spam. AI systems are mostly rule based scoring system. They associate different scores with different keywords depending on the criteria. If those keywords occur within the email's header or content, their associated scores are summed up. If the overall score of the email is greater than a minimum threshold value, email is declared as spam. Spam Assassin [20] uses this rule based scoring system.

Research has also been done for automated filter construction as well. Sahami et.al [21] used a decision theoretic framework and probabilistic learning methods. He was able to produce very accurate filters by exploiting domain-specific features along with email's raw text. Reactive spam filtering techniques to fight spam also exist. In these techniques the users of the email system report the messages as spam or legitimate. The reputation of the reporter is very important in this system and a report from a user with high reputation is more authentic as compared to the one from a less reputed reporter. Zheleva [22] have proposed spam filtering systems which rely on reputation of the reporter.

Kong et.al [23] suggested a distributed spam-filtering system. This filter uses properties of the social networks. Similarly Boykin et.al [24] utilized the social networks to fight against the spam. According to them the social networks are useful enough to judge the trustworthiness of outsiders. Chirita et.al [25] utilized a ranking algorithm for the email senders. They have suggested discrimination of the messages on the bases of the MailRank score of the sender. Other techniques to fight against email spam being used include Maximum Entropy Model [26] and Memory Based Learning [27].

2.1.3 Spamming in Blog

Spamming in blogs also known as link spamming or comment spam is a kind of Spam, which is initiated to achieve top rankings for search engines. This is used to spread links while commenting on the others blogs, forums and other web pages. In a research it is determined that around 83% of all comments are spam [28] which creates problems for the whole blogging community. Spammers visit blogs and leave comments containing advertisements and links pointing back to their own websites. These comments do not add much to the knowledge base of the blog. Mostly the comments contain material which is not related to

the contents of the blog post. Spammers also use spam to promote the commercial products. Comments containing pornographic links and other offensive material are also posted on the blogs. The spammers try to distract the visitors of the blog and encourage them to visit spammer's website. This helps the spammers to gain higher search engine ranking and higher traffic for their websites. This spoils back-linking frequency dependant search engine rankings. Today spammers use automated software programs to find blogs and such websites where they can post comments containing links and advertisements to other sites and products.

2.1.4 Methods to Fight against Link Spam

Blogging software development firms and researchers are using different methods to fight against spamming. These methods include:

2.1.4.1 CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) is a program that protects websites against bots. It generates and grades tests that humans can pass but current computer programs cannot. Many different types of CAPTCHA are present today. Major categories of CAPTCHA include mathematical, simple, and symbolic. Now a days Captcha.com [29] has provided a new type of CAPTCHA called re-captcha in which user needs to type in two distorted words instead of one as was the case with its ancestors. Figure 2.1 shows a re-cpatcha example. The words cannot be read by computer programs but a human can recognize them with ease.

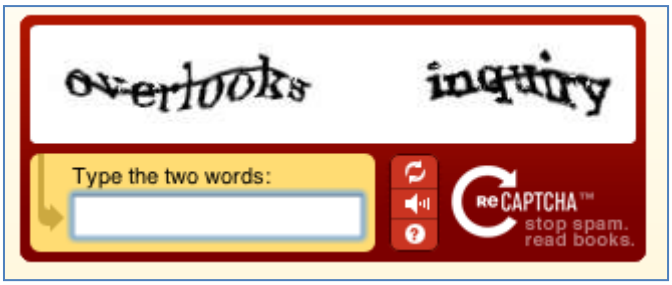


Figure 2.1: Recaptcha from http://www.captcha.net/

2.1.4.2 Reject Continuous Submissions

Mostly the spam comments are posted in bulk by the computer software programs that is normally associated with the term flooding. Although we cannot stop this spamming act completely, but we can minimize it by rejecting the continuous stream of submission of comment form. In this way the computer program will be able to post the link spam for very few numbers and rest will be rejected by our software.

2.1.4.3 Human User Moderated Comments

Humans can recognize the comment spam very easily. If a human has enough time for checking the nature of all comments, this will almost eliminate the comment spam problem once for all. But this is a very time and effort consuming task for humans and still there are chances of human errors. Thus it is a costly and still not robust enough.

2.1.4.4 Don't Allow URL in Comment

We can stop the link spam by disallowing postings of comment having URLs and hyperlinks in it. In this way we will be helping the search engines but still the spam is there on our blog and this spam will be spoiling our original user comments and posts.

2.1.4.5 Use REL= 'NO FOLLOW'

A new attribute named REL was introduced by Matt Cutts from Google and Jason Shellen from Blogger in 2005. The REL attribute specifies the relationship of the linked document with the current document. One can intimate the search engines not to spoil their ranks by including the link in their directory. This can be done by adding this REL attribute with value 'No Follow' in all the URLs pointing to other sites from our blogs. Google, MSN and Yahoo regard this attribute and do not count the link as a valid link for Search Engine Optimization (SEO). Again this is helpful to the search engine crawlers not to include the links to their directories but still the blogs will be facing the spam problem. The humans will have to clean these comments manually.

2.2 Comment Text Analysis

Researchers have been working on content analysis to fight against the link spam and they take into account the number of links and presence of spam words in the comment. They take care of the blacklisted IP addresses as well. If the comment is coming from a blacklisted IP address then the system should not accept it. Wu et.al [31] presented a directed approach to extracting link spam communities. Their method starts with a small user provided spam seed set. They simulate a random walk on the web graph. They have used decay probabilities for random walks to explore the neighborhood around the seed set. Truncation is used to retain only the most frequently visited nodes. The nodes are sorted in decreasing order of their final probabilities and presented to the user. Zhou et.al [32] has used page farms to detect link spam. They have suggested two spamicity measures based on page farms. They can be used as an effective measure to check whether the pages are link spam target pages. Gyongyi et.al [33] introduces the concept of spam mass, a measure of the impact of link spamming on a

page's ranking. It is also recommended that only registered users should be able to post and moderate the comments on blog and where as non-registered users can only post the comments. Registered users then should be involved to moderate those comments before they could be published for the public. All these efforts are helpful in minimizing the link spam but none of them can guarantee 100 percent for stoppage of comment spam.

2.3 Spamming on Twitter and Other Micro-blogging Platforms

Most of the users are sending spam messages on micro-blogs also. Twitter is selected in micro-blogging platforms. . Spammers are wasting Twitter's resources and creating problems for its users by polluting trending topics with their orthogonal tweets. Twitter has been targeted by companies for promoting their brands such as #MOONFRUIT campaign. Its promotions containing training courses and even some messages also contain pornographic links. The problem of spam detection on micro-blogging platforms is more difficult as compared to the macro-blogging platforms because micro-blogs allows a very limited number of characters and detecting spam in small chunks of text is even more difficult. The public message on micro-blogs such as advertisements, are considered as spam by one user but not by the others due to the diversity in their interests. The users who are sending legitimate messages most of the times and occasionally sending spam message cannot be adjudged as spammer unlike the macro-blogging and emailing platforms where we can declare the users as spammers and non-spammers. Many researchers are working on spam detection in short text messages. Cormack [30] addressed the issue of content-based spam filtering for short text messages. These short messages can be from mobile (SMS) communication, blog comments, and abridged email information to be shown on a low-bandwidth client. These messages normally contain few words. According to them these

messages are challenges to bag-of-words based spam filters. They have concluded that compression-model filters perform quite well on small chunks of text.

2.4 Related Work

Opinion mining, sentiment analysis, spam detection and social networking phenomenon on Twitter have emerged as a field of interest for many researchers. This section presents work done in the vicinity of Twitter and other micro-blogging platforms.

Chris Grier et.al [46] analyzed the unique features of Twitter being utilized by spammers to spread links and many other ways. Their work shows that Twitter is a highly successful platform for driving users to visit spam with a click through rate of 0.13% which is much higher than the same things on emailing platform. They have analyzed about 25 million URLs and found that 8% of these URLs landed on phishing, malware, and scams listed on popular blacklists.

Akshay Java et.al [47] studied the topological and geographical features of Twitter's social network. They have concluded that most of the people use micro-blogging for either seeking or sharing information and some users talk about their daily life activities. They have also given an idea about the connection between users with same interests. According to them the major intentions of the users on Twitter include daily chatting, conversations, information sharing, reporting news, friends and information seekers.

According to Alexander Pak et.al [48] millions of users on Twitter communicate their daily life activities by posting opinions on different aspects of life. This research performed linguistic analysis of tweet content. Authors have built a sentiment classifier which can determine positive, negative and neutral sentiments in English language tweets.

Bharath Sriram et.al [49] provides evidence about the overwhelming of the Twitter users with raw data. They have suggested the classification of the tweets into different categories. They have extracted domain specific features from Twitter user profiles and then utilized these features to classify the data into major categories such as news, events, opinions, deals and private messages.

2.5 Critical Analysis

Although many different algorithms do exist today to detect spam messages from emails and link spam on blogs and blogging platforms, but these algorithms are not very handy when problem of spam detection is addressed at the micro-blogging level and social networking platforms due to the short length of text message. The work already done to measure similarity between short text chunks was utilized to give an idea about the nature of the short text messages. Measuring the similarity among the different parts of the tweet (a short text message form Twitter) can help to make a decision about the tweet. For this purpose some of the research work on short text message analysis was critically reviewed,. in which researcher have tried to measure the similarity between two short text chunks.

Gabrilovich et.al [11] utilized Wikipedia to expand the short texts and then measured the similarity between these short texts. This method proves to be equally good for short and large texts. They also provided word sense disambiguation by exploring the context of the neighbors of the word. This method is unable to handle new terms and requires pre-processing of Wikipedia articles.

Christopher Meek et.al [12] used search engines for expanding short texts and then measured the similarity between the text chunks in the expanded universe. Their method proves to be better than the previous one as it provides coverage for the new terms and also does not need

to pre-process a large dataset. This method provides mechanism to fine tune the learning algorithms according to the needs to the target application, but this has two drawbacks also. One is the use of the non-standard dataset to measure the algorithm performance and second is the amount of noise produced due to expansion of the original text with the help of search engines.

Kanaris et.al [13] introduced content based spam detection which uses character level N-grams. They have used Support Vector Machines (SVM) algorithm. This approach does not need usage of any lemmatizer or text pre-processing. Although they have used Ling Spam - a standard dataset for evaluation and have achieved good results on it, however this fact cannot be ignored that they have used lower case copy of this standard dataset. They only evaluated 3, 4 and 5 character grams. Due to the character grams a very large number of attributes is introduced due to which the choice of the machine learning algorithms becomes limited to only support vector machines.

CHAPTER 3

METHODOLOGY

3.1 Overview

The goal of the research is to design an algorithm through which we can detect spam in short text-based messages on micro-blogging platforms such as Twitter. As discussed earlier in Chapter 1, spammers are exploiting Twitter to spread spam messages. We have devised a generic algorithm by applying semantic analysis and similarity measures among different parts of the tweets to detect spam.

We collect the details about each hashtag and title of each URL, expand them using search engine result snippets (a technique similar to query expansion), then apply Natural Language Processing (NLP) and semantic analysis on these expanded search results to extract semantic information. This semantic information is utilized in the similarity measures to find the similarity score among different parts of the tweet. These similarity scores contribute towards the spam index of each tweet, along with tweet author's ranking and spam to legitimate factor. The final spam index is used to make the final decision about the status of the tweet for that particular user.

The chapter has been divided into six sections. Section 3.1 describes different operations performed on a tweet, including comparison of the tweet text with a list of spam words maintained by the user, extraction of different components of tweets including URLs, hashtags, user mentions and re-tweets and finally cleaning the tweet by removing these components from it. Section 3.2 elucidates the text expansion process utilizing search engine result snippets. Semantic analysis and information extraction are discussed in Section 3.3. Section 3.4 describes the similarity measures among different parts of the tweet. Section 3.5 introduces the tweet author rank and his spam to legitimate tweet ratio. In Section 3.6 we

have explained the algorithm designed for calculating the spam index for the tweet and making the final decision about the tweet status for a particular user after making a comparison with the user mentioned spam threshold value.

3.2 Tweet Operations & Extractions

After collecting the tweet from Twitter via its API, extraction processes are performed to determine different components of the tweet. These components are removed from the tweet to obtain cleaned raw tweet text. Text expansion via search engines, semantic analysis and similarity measures are performed on these components of the tweet and its raw text to find its spam index and hence, make the final decision about the status of the tweet. These steps are described below.

3.2.1 Hashtag Extraction & Hashtag Details Collection

As discussed in Chapter 1, the Twitter community has adopted a mechanism for adding context and metadata to the tweets via hashtags. A tweet can have zero, one or more hashtags. All hashtags are extracted from tweets and the record of these saved in a repository for further usage during the later stages of the research.

3.2.2 URL Extraction, Reversal and Title Collection

A tweet can have zero, one or more URLs. Spammers are spreading their website URLs in a flimsy way by utilizing the URL shortening exploits. To cope with this issue we extract URLs from the tweet and expand them for the users. There are many advantages of this approach such as the first hand knowledge to the user about the destination of the URL without any click. The second advantage from this is the contribution of the URLs towards the spam score calculation of the tweet. The extracted URLs are reversed and followed by our

program. These URLs and their titles are extracted and stored in repository with a link back to the tweets containing them.

3.2.3 RT & User Mention Extraction

Twitter provides the facility to send direct message to other users by including @UserName anywhere in the tweet. Users can even repost the messages sent by other users by just adding RT: @UserName in the start of the tweet. This phenomenon is called re-tweeting. Such user mentions are extracted from the tweets and stored in the repository for further usage in spam score calculation algorithm.

3.2.4 Tweet Cleaning and Raw Text Extraction

As discussed earlier a tweet may contain hashtags, user mentions and URLs along with raw text. In this step we remove all the occurrences of these components from the tweet. In this way we get raw tweet text. This raw text is later expanded using our expansion algorithm in order to measure the spam score of the tweet.

3.3 Text Expansion

Missing context and background information problems arise while dealing with the short text. In order to cope with these issues, we have utilized text expansion mechanism. The short text is expanded by conducting a search for this short text against information enriched sources, such as search engines, domain specific directories and Wikipedia [35]. Search result snippets thus obtained are combined together to construct a longer text chunk. This provides better background knowledge and context about the short text in question. This helps to find the facts and figures from the short text. Named Entities (NEs) hidden within the short text are obtained with better understanding as well.

Some researchers have utilized only Wikipedia [35] as information resource but we face a problem while we dealing with new terms and text snippets. Wikipedia alone cannot provide us an adequate amount of search results relating to these new terms. Search engines are best known for providing a fair amount of results to expand a given text. For this reason, search engine like Google [36] are preferred for text expansion. Text expansion is performed separately on tweet raw text, each URL title and hashtag details.

3.4 Information Extraction & Semantic Analysis

We perform named entity recognition to extract events, facts and figures present in the short unstructured text. Through semantic analysis these NE, events, facts and figures are linked with an external resource, which helps to disambiguate the information as well. As an example consider a very simple sentence: "Philip Sanford is new CEO of Jackson Hewitt Inc.". Information extraction will retrieve two NEs - a person "Philips Sanford" and an organization "Hewitt Inc." and one function, a CEO. Semantic analysis on this can not only extract relationship between "Philip Jackson" but can also link the extracted entities with an external knowledge base such as DBpedia [37]. The most important value addition in semantic analysis is the ability to disambiguate entities. At the end, it will be possible to link the concerned piece of text to relevant external information.

Information extraction and semantic analysis tools normally use NLP as the core technology. Some powerful web services are also available to perform semantic analysis of text. e.g. OpenCalais [38] and OpenAmplify [39]. These services provide information about the text, NEs present in the text, relationship, events and other facts hidden within the text. These services also attach a confidence measure to the extracted pieces of information.

The semantic analysis is performed on the expanded text and we store all the resulting NEs, events, facts and figures in the repository for further processing. Algorithm described in Figure 3.1 is designed for expanding the text using search engine results and then, extracting semantic information from this expanded text.

Purpose:	Expands input text *t* using search engine results and extracts semantic information from the expanded text.
Input:	Text *t*
Output:	List of named entities *N*, List of relationship *r*, List of facts & figures *F*, List of events *E*

Steps:

1. Search *t* on a search engine of your choice
2. For each result snippet
    a. Combine title and short description to get *td*
3. Combine all *td* collected from step 2 to get an expanded text chunk *T1*
4. Pass text *T1* to your semantic analysis software
5. Extract Named Entities and store in list *N*
6. Extract relationship among named entities and prepare a list *R*
7. Extract Events and store in a list *E*
8. Extract facts and figures and store in a list *F*
9. Return *N,R,E and F*

Figure 3.1: Text Expansion & Semantic Information Extraction Algorithm

3.5 Measuring Similarity

Similarity measures are carried out on the semantic information collected from different text chunks - extracted and expanded in the above steps. As discussed in previous sections a tweet contains different components such as URLs, hashtags and some raw text. In order to measure similarity among these components, text expansion algorithm described in Figure 3.1 has been utilized. Similarity among these components is measured after expanding and extracting semantic information from these texts. Inter component and intra component similarity is measured. Following possible combinations can be used for this purpose.

- Amongs URLs
- Among Hashtags
- Hashtag to URL
- Hashtag to raw tweet text
- URL to raw tweet text

These similarities are averaged out to make a combined effect in the spam detection algorithm defined in Section 3.6

3.6 Spam Detection

On Twitter it is relatively hard to declare a user as spammer unlike email or blog. This is one of the major challenges. The main reason is the verity that the author of a spam tweet might have been sending legitimate tweets most of the times and an occasional spam message. To cope with this challenge, we have taken into consideration the ranking and spam to legitimate tweet ratio for tweet authors. Factors for this ranking may include user's total tweets, number of re-tweets, number of followers, number of friends, tweet timings and tweet frequency etc.

To make final decision about the status of a tweet we have developed the algorithm described in Figure 3.3. Major factors in the algorithm are the average similarity measure (among different parts of the tweet), author's spam to legitimate tweet percentage and spam words list maintained by the user (for which we need to make a decision about the tweet).

Purpose:	*Calculates tweet spam index*
Input:	Tweet text T, Average of similarities(among tweet components) A, Author's spam to legitimate percentage R, Threshold H, Spam words List L
Output:	Tweet Status (either spam or legitimate)

Steps:

1. For each word W in T:
    a. Search W in L.
        i. If W is found in L Then
            1. Declare T as spam
            2. Exit
        ii. If W is not found in L Then
            1. Go to Step 2
2. Compare A and H
    a. If A < H Then
        i. Declare the tweet as legitimate
        ii. Exit
    b. ELSE If A ≥ H and R <25 Then
        a. Declare T as Legitimate
        b. Exit
    c. If A ≥ H and R ≥ 25 Then
        a. Declare T as Spam
        b. Exit

Figure 3.2: Spam Detection Algorithm

CHAPTER 4

IMPLEMENTATION & RESULTS

This chapter describes the implementation details of our work, tests performed and results achieved. Section 4.1 covers APIs including Twitter API for collecting tweets, Five Filters API for term extraction, Google search API for expanding short text and Open Calais for extracting semantic information from the given text segments. Section 4.2 describes similarity measure – Jaccard co-efficient. Section 4.3 discusses the evaluation of our work and concludes with results achieved. A discussion on system variables is presented in Section 4.4. These system variables can be tuned to achieve maximum accuracy.

4.1 Application Programming Interface (APIs) Used

The APIs used in our work include Twitter API for accessing Twitter data including the meta data about a tweet and tweet author as well. Five Filters term extraction API extracts the keywords and terms present within a text chunk. Google search API provides term expansion facilities while Open Calais API has been utilized as semantic information extraction engine.

Following sections describe the implementation details of these APIs along with sample inputs and outputs.

4.1.1 Twitter API

Twitter provides a restful API [41] to access its data. This API provides data about authors & tweets in Extensible Markup Language (XML) and JavaScript Object Notation (JSON) formats. This API provides powerful search capabilities as well. We used this API to collect tweets, authors of the tweets, social structure of these authors and other metadata including tweet publishing time etc. We stored all this information in a relational database for further processing in the next steps of the algorithms designed and implemented in our work. We

utilized JSON format while collecting data from Twitter. Figure 4.1 presents a sample response from Twitter API.


```
Array
([results] => Array
    ( [0] => Array
        ([profile_image_url] => http://a1.twimg.com/profile_images/1104751437/N_normal.jpg
            [created_at] => Thu, 23 Sep 2010 06:06:36 +0000
            [from_user] => Naushi_Ifti
            [metadata] => Array
              ([result_type] => recent)
            [to_user_id] => [text] => RT @TweetMidget: http://tinyurl.com/CWF-FemCreative Prospects for a Female #Copywriter #Pakistan #Advertising #Marketing #Brandbuilding #ProfessionalIssues
            [id] => 25283411404
            [from_user_id] => 149640097
            [geo] =>
            [iso_language_code] => en
            [source] => <a href="http://Twitter.com/">web</a>
  [max_id] => 25283411404
  [since_id] => 0
  [refresh_url] => ?since_id=25283411404&q=%23Pakistan
  [next_page] => ?page=2&max_id=25283411404&rpp=1&q=%23Pakistan
  [results_per_page] => 1
  [page] => 1
  [completed_in] => 0.019382
  [query] => %23Pakistan
```


Figure 4.1: Twitter Search API response

4.1.2 Five Filters Term Extraction API

Term extraction is a subtask of information extraction. The goal of term extraction is to automatically find significant terms from a given corpus. Term extraction is a helpful in measuring semantic similarity, knowledge management, human and machine translation, etc. We used Five Filters API [42] to extract the terms from our original text chunks for tweet, URL details and hashtag details. These terms are used as input to our similarity calculation algorithm. We extracted the terms from the input texts and stored them in our repository for further processing during the similarity measuring steps in our work. A sample term extraction output from this API is shown in Figure 4.2.

Input Text	**Extracted Terms**
Inevitably, then, corporations do not restrict themselves merely to the arena of economics. Rather, as John Dewey observed, "politics is the shadow cast on society by big business". Over decades, corporations have worked together to ensure that the choices offered by 'representative democracy' all represent their greed for maximised profits[50]	1. Arena 2. Business 4. Cast 5. Choices 6. Corporations 7. Decades 8. Democracy 9. Dewey 10. Economics 11. Greed 12. John 13. John Dewey 14. Maximised 15. Maximised Profits

	16. Politics 17. Profits 18. Representative 19. Representative Democracy 20. Shadow 21. Shadow Cast 22. Society

Figure 4.2: Five Filters Term Extraction API sample result

4.1.3 Google Search API for Term Expansion

In order to explore the terms further for extracting semantic information from tweet text, we apply term expansion. Google Search API [43] is used for this purpose. We extract title and summary of first eight search results as suggested by Google search engine against each term. Figure 4.3 shows a tweet text segment and its expanded paragraph.

Short Text	**Expanded Text**
Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative!	Macbook Pro Giveaway - Moonfruit - Beautiful websites, simply Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC #Moonfruit. Tweet this! Follow @moontweet ... Twitter censors Moonfruit? What does it mean for the future of ... Jul 6, 2009 ... Important Update â€“ Big Bang Finale! The competition response has been crazy and wonderful. ... Now back to hoping to win a MacBook, LOL #Moonfruit ..... over the past 10 years, and continuing to work in the industry, ... Randy Sacchetti (ransac7) on Twitter Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative!

	http://bit.ly/96bxC #Moonfruit 1:27 AM Jul 7th, ... Jorge Beltran (JorgeBeltran) on Twitter Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC #Moonfruit 9:28 AM Jul 7th, 2009 via web ... Best of Liverpool supplement August 2010 To reflect 2010's status as Year of Health and Wellbeing, throughout the summer ... LIVERPOOL'S Bang on Top Productions, creators of hit comedy plays One Night in Istanbul ..... Visit: www.alicelenkiewicz.Moonfruit.com/. BEST OF LIVERPOOL .... Herzegovina in the final home game of the EuroBasket Qualifying Rounds. ... popurlsÂ® | archive | issue 09-06-30 Jun 30, 2009... Macbook Pro Giveaway - Moonfruit - Beautiful websites, simply ..... The media covers the final act of Michael Jackson's death by reporting that the wall -to-wall media coverage... Mother of the year candidate leaves infant home alone to go ..... Bing, Bang, Boom... And a zap. - Diggnation ... Warung Senggol :: Featured Podcast :: Bincang Santai Warung ... Feb 21, 2009 ... lang> :<Twitter enlang> :</Twitter : #Moonfruit sounds like ... lang> :<Twitter enlang> :</Twitter : #mw2 possible or not final ..... lang> :<Twitter en lang> :</Twitter : @iamRE reeeeeeeeeeee! why u gotta bang on me? ..... lang> :<Twitter enlang> :</Twitter : @jenferjenfer nah-I love my macbook.

Figure 4.3: Text Expansion via Google Search API

4.1.4 Open Calais API for Semantic Information Extraction

In order to perform semantic analysis we need entities, events and actions presented in the expanded text. For this purpose expanded text was passed to Open Calais [38] – a semantic information extraction service. This information returned by API includes categorized NEs, relationship among entities, events, actions, social tags etc. The API associates a confidence value with every piece of information extracted from the input text. The confidence value is between 0 and 1 and represents the degree of belief of the API about the extracted information. This value is used in measuring the similarity among different text segments. Sample output of this API is depicted in Figure 4.4.

Input	Output
Macbook Pro Giveaway - Moonfruit - Beautiful websites, simply Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC #Moonfruit. Tweet this! Follow @moontweet ... Twitter censors Moonfruit? What does it mean for the future of ... Jul 6, 2009 ... Important Update â€“ Big Bang Finale! The competition response has been crazy and wonderful. ... Now back to hoping to win a MacBook, LOL #Moonfruit ..... over the past 10 years, and continuing to work in the industry, ... Randy Sacchetti (ransac7) on Twitter Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC #Moonfruit 1:27 AM Jul 7th, ... Jorge Beltran (JorgeBeltran) on Twitter Big bang finale! Celebrate 10 years of Moonfruit and win a Macbook Pro. Be creative! http://bit.ly/96bxC #Moonfruit 9:28 AM Jul 7th, 2009 via web ... Best of Liverpool supplement August 2010 To reflect 2010's status as Year of Health and Wellbeing, throughout the summer ... LIVERPOOL'S Bang on Top Productions, creators of hit comedy plays One Night in Istanbul ..... Visit: www.alicelenkiewicz.Moonfruit.com/. BEST OF	**Social Tags:** 1. Apple Inc. 2. World Wide Web 3. Computing 4. Big Bang 5. Macintosh 6. MacBook 7. Moonfruit 8. Twitter 9. Macbook Pro 10. Personal computers 11. MacBook family **Entities**: • **City:** o Istanbul,Turkey o Liverpool,England,United Kingdom • **Company** o Twitter Inc. • **Industry Term** o to-wall media coverage • **Movie** o One Night

LIVERPOOL .... Herzegovina in the final home game of the EuroBasket Qualifying Rounds. ... popurlsÂ® | archive | issue 09-06-30 Jun 30, 2009... Macbook Pro Giveaway - Moonfruit - Beautiful websites, simply ..... The media covers the final act of Michael Jackson's death by reporting that the wall -to-wall media coverage... Mother of the year candidate leaves infant home alone to go ..... Bing, Bang, Boom... And a zap. - Diggnation ... Warung Senggol :: Featured Podcast :: Bincang Santai Warung ... Feb 21, 2009 ... lang> :<Twitter enlang> :</Twitter : #Moonfruit sounds like ... lang> :<Twitter enlang> :</Twitter : #mw2 possible or not final ..... lang> :<Twitter en lang> :</Twitter : @iamRE reeeeeeeeeeee! why u gotta bang on me? ..... lang> :<Twitter enlang> :</Twitter : @jenferjenfer nah-I love my macbook.	• **Person**
	  ○ Jorge Beltran
	  ○ Michael Jackson
	  ○ Randy Sacchetti
	• **URL**
	  ○ http://bit.ly/96bxC
	  ○ www.alicelenkiewicz. Moonfruit.com
	Events & Facts:
	• Generic Relations
	  ○ Twitter Inc., Moonfruit, censor
	  ○ creators of hit comedy, One Night, play

Figure 4.4: Semantic Information Extraction via Open Calais API

4.1.5 Twitter Grader API for Tweet Author Ranking

As discussed in the previous chapters, one of the important factors in our algorithms is the author's trustworthiness. If an author sends legitimate tweets most of the times with an occasional spam, we cannot adjudge the author as a spammer and vice versa. To make a sensible decision about author's credibility, we are using Twitter Grader API [40]. This API

allows checking the power of a Twitter user as compared to the other users. It measures the influence and power of a Twitter user by associating a rank and grade with the user. This ranking is based on different factors including:

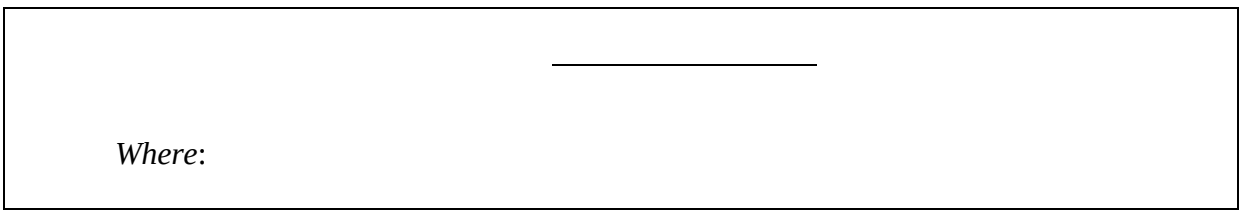
- Number of followers
- Power of followers
- Number of tweets
- Follower/Following Ratio
- Engagement

4.2 Similarity Measure

We have used Jaccard co-efficient to measure similarity among the text segments. It helped to determine overlap between two text segments. Different pieces of the information (extracted from original text or its expanded version) are used as binary attributes for Jaccard coefficient. Each of these pieces is either present in any one of the text segments or in both segments. The attributes being used in our work include:

- Terms extracted from original text
- Social tags suggested by our semantic information extraction engine
- All NEs, events, facts and figures with a confidence measure of greater than or equal 33%

Formula to measure Jaccard co-efficient is given below:



Where:

M_{11}	=	A set of attributes present in both texts
M_{01}	=	A set of attributes present in second text only
M_{10}	=	A set of attributes present in first text only

Figure 4.5: Jaccard co-efficient Formula

4.3 Evaluation

This section consists of three parts including evaluation methods, description of the data set and finally the experiments performed and results achieved by our algorithm.

4.3.1 Evaluation Methods

In order to evaluate the system, recall and precision measures have been used. Recall is a measure of completeness whereas precision is measure of the accuracy. They evaluate the quality of an un-ordered set of retrieved items. Formulae for recall and precision are given below:

4.3.2 Dataset

A dataset of about 40,606 tweets was collected from Twitter. All the tweets are relevant to a marketing campaign run by Moonfruit on the occasion of their 10^{th} anniversary in July 2009. The tweets were collected from July 07, 2009 to July 18, 2009. These tweets were sent by 512 different users. 1,691 tweets from this dataset are manually labeled as spam or legitimate

by experts. 1,397 of them are labeled as spam and 294 as legitimate. Table 4.1 shows some key statistics of the dataset:

Feature	**Value**
Total Tweets	1,691
Spam Tweets	1,397
Legitimate Tweets	294
Tweet contributing users	512
Tweet Collection Time	30 May 2009 to 8 August 2009

Table 4.1: Dataset Statistics

Table 4.2 shows the statistics of the extracted information from the dataset.

Feature	**Value**
Total URL Mentions	273
Unique URL Mentions	51
Total Hashtag Mentions	2,729
Unique Hashtag Mentions	512
Total Named Entities Mentions	38,540
Unique Named Entities Mentions	7,454
Total Categories of Named Entities	39

Table 4.2: Extracted Information Statistics

4.3.3 Experiments

System variables involved in our experiments include confidence measure (given by the semantic information extraction system about an entity) and threshold value for spam set by the user. These two variables were initialized with 0.1. Threshold was incremented with 0.1 until it reached to the value of 1.0 and then afterwards confidence measure was incremented with 0.1 and threshold was reset back to 0.1. Thus in total 100 experiments were performed for different values of confidence measure and threshold. Dataset described in Section 4.3.2 was used to evaluate the spam detection algorithm. As described in previous section the evaluation methods used to validate our results are recall, precision and F-measure. Top 5 results obtained from these experiments are presented in the Table 4.2.

Serial Number	Threshold	Confidence	Precision	Recall	F-Measure
1	0.2	0.3	0.97	0.94	0.95
2	0.3	0.6	0.94	1.0	0.97
3	0.5	0.5	0.92	1.0	0.96
4	0.6	0.7	0.82	1.0	0.90
5	0.9	0.8	0.81	1.0	0.90

Table 4.3: Results

Algorithm achieved a precision of 97% with a recall of 0.94 and F-measure of 0.95. These values were achieved at a threshold of 30% and a confidence measure of 60%. In another set of experiments with different values set for confidence and threshold. We were able to achieve a precision of 94% and a recall of 100% as well. In this set algorithm obtained an F-measure of 97%.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

In previous chapters we have discussed semantic analysis of short text messages on micro-blogging platforms specifically for Twitter. Different aspects of the short text messages have been explored. The most important aspects include the detection of the relationship among different parts of the messages. An algorithm has been developed for spam detection from these messages. Semantic information (extracted from the search engine based expansion of the short text messages) is used in the algorithm. The similarity among different parts of short text has been measured with the help of the semantic analysis. The average similarity among different parts of the message is compared with a user defined threshold to make a final decision about the status of the tweet. To evaluate and validate the algorithm, it was applied on Twitter dataset collected through their API.

This chapter summarizes the contributions, discusses various applications of the spam detection algorithm and the possible future extensions of our work. These include usage of the machine learning techniques, plugins for Wordpress[8] and other popular blogging and short messaging services and tools.

5.1 Conclusions

In present age, communication over the internet has become inevitable. People are using emails, blogs, social networks and online discussion forums to communicate with their peers and communities worldwide. Introduction of micro-blogging tools like Twitter have provided a new dimension to share ideas, products and other important pieces of information. One of the biggest evils being faced by online communication channels is spamming. Today people receive too many un-necessary messages including new product promotions, advertisements,

[8] http://www.wordpress.org

discounted sales offers and lot many things like this. This spoils server bandwidth, internet resources and most importantly internet users' precious time. Dealing with short text messages (over the micro-blogging platforms) poses new challenges. e.g. One message can be considered spam by one user but legitimate by others due to the diverse difference in their interests. Similarly the sender of short messages cannot be declared spammer. Thus spam detection on micro-blogging platforms adds new dimensions to the challenges of spam detection.

The major focus of this research is the development of the algorithm to detect spam in the short text message. Semantic analysis on short text messages in an expanded universe (achieved from Google search engine results) has been utilized in this algorithm. Similarity among different parts of short message has been calculated and then compared with a user defined threshold to make a final decision about the status of the short text message. The algorithm has been tested on Twitter short messages and generated fantastic results. A precision of 97% was achieved while keeping the recall and F-measure at equally good levels of 94% and 95% respectively.

5.2 Contributions

5.2.1 Semantic Analysis of Tweet Content

Semantic analysis has been performed on short text message from Twitter. Named entity recognition, disambiguation and categorization has been provided by Open Calais API [38]. In this process system was able to recognize named entities, relationship among them, events and facts from the tweets text. Thus with the help of semantic analysis we were able to recognize the most commonly used URLs, hashtags, highly referred users, mostly referred named entities and their categorization.

5.2.2 Spam Tweet Detection Algorithm

An algorithm for spam tweet detection has been designed and implemented in this research. This algorithm detects spam out of short text messages on micro-blogging platforms. This algorithm used semantic analysis described in Section 5.2.1. Evaluation of this algorithm has been done using precision, recall and F-measure. The algorithm results are depicted in Table 4.2.

5.2.3 Detecting Social Networking Relationships among Users

Semantic information extraction has provided us a good opportunity to do some other challenging works with the short text messages. We have utilized semantic information to recognize social networking relationships among users on Twitter. Self organizing maps and night sky visualization algorithms have been applied on semantic information. We detected the relationships among Twitter users without knowing their follower/following relationships. We used the inclination of the users towards usage of Hashtags, URLs, user mention and named entities in their tweets. We are in process of writing a research paper on this work as well.

5.3 Possible Applications

As discussed earlier spam has become a gigantic problem on almost all possible communication channels over the internet. Although we have developed the algorithm to deal with this issue on the micro-blogging platform yet this algorithm can be extended and applied to many other domains for the same purpose. Some prominent areas where our algorithm can be applied include:

- Social networking and other micro-blogging and platforms like Facebook[9], Google Buzz [10], LinkedIn [11]and YouTube[12] etc.
- Discussion forums for comment spam
- Blogs for comment spam
- Ecommerce Websites - product reviews by customers

5.4 Future Direction

We are planning to build a browser based plug-in for Twitter users. The plug-in will be based upon our spam detection algorithm. This plug-in will help users detect spam from a set of tweets being displayed on the Twitter website. Users will have facility to tune the system according to their preferences. They will be able to modify following factors according to their need:

1. Weights of similarity among different parts of a tweet.
2. Threshold value for spam index to declare a tweet as a spam
3. Confidence measure of the semantic information with which Open Calais makes a decision about a piece of information

[9] http://www.facebook.com
[10] http://www.google.com/buzz
[11] http://www.linkedin.com
[12] http://www.youtube.com

REFERENCES

[1]. Twitter http://www.Twitter.com as on 15 June 2009

[2]. Nielsen: Twitter Was Fastest Growing Community Last Month http://www.readwriteweb.com/archives/nielsen_Twitter_was_fasting_growing_community_last_month.php as of 28 January 2010

[3]. Dell, Moonfruit Claim Twitter Campaigns Effective http://www.marketingcharts.com/topics/branding/dell-Moonfruit-claim-Twitter-campaigns-effective-10198/ as on 7 November 2009.

[4]. Tweeting is more than just self-expression http://live.psu.edu/story/41446 as on 25 April 2011.

[5]. Free Website Builder - Moonfruit - Total website design control http://www.Moonfruit.com/macbook-pro.html as on 25 April 2011

[6]. What is Moonfruit? A Twitter Campaign | Scoop News http://www.scoop.co.nz/stories/BU0907/S00287.htm as on 25 April 2011

[7]. Xin Li, Bing Liu and Philip S. Yu, "Mining Community Structure of Named Entities from Web Pages and Blogs", American Association for Artificial Intelligence (AAAI), 2006

[8]. Takaaki Hasegawa, Satoshi Sekine and Ralph Grishman, "Discovering Relations among Named Entities From Large Corpora", ACM, 2004

[9]. Tim Berners Lee, James Hendler & Ora Lassila, "The Semantic Web", Scientific American, May 2001

[10]. Twitter Grader | Get Your Twitter Ranking http://Twitter.grader.com/ as on 10 February 2010

[11]. Evgeniy Gabrilovich , Shaul Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", 20th International Joint Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI), 2007

[12]. Wen-tau Yih and Christopher Meek, " Improving Similarity Measures for short segments of text", Association for the Advancement of Artificial Intelligence (AAAI), 2007

[13]. Ioannis Kanaris, Konstantinos Kanaris and Eftathios Stamatos, "Spam Detection Using Character N-Grams", In Proc: Proceedings of the 4th Hellenic Conference on AI (SETN 2006), Springer LNCS, 3955, pp. 95–104, 2006.

[14]. Internet Growth Statistics – Global Village Online http://www.internetworldstats.com/emarketing.htm as on 13 March 2010

[15]. MAPS - Support - Definition of Spam http://www.mail-abuse.com/spam_def.html as on 13 March 2010

[16]. Analysis of Spam by Anselm Lambert http://en.scientificcommons.org/17540011 as on 15 March 2010

[17]. Review-DATA-MINING-CUP http://www.data-mining-cup.com/2003/Wettbewerb/1059704704/ as on 15 March 2010

[18]. Jos'e Maria Go'mez Hidalgo, "Evaluating cost-sensitive Unsolicited Bulk Email Categorization,", ACM symposium on Applied computing, ACM, 2002

[19]. Harris Drucker, Donghui Wu, Vladimir N. Vapnik, "Support vector machines for spam categorization", IEEE Transactions on Neural Networks, IEEE, 1999, vol. 10. pp: 1048-1054

[20]. SpamAssassin: Welcome to SpamAssassin http://spamassassin.apache.org/ as on 25 April 2011

[21]. Mehran Sahami and Susan Dumais and David Heckerman and Eric Horvitz, "A Bayesian Approach to Filtering Junk E-Mail In Learning for Text Categorization", Association for the Advancement of Artificial Intelligence (AAAI), 1998

[22]. Elena Zheleva and Aleksander Kolcz and Lise Getoor, "Trusting spam reporters: A reporter-based reputation system for email filtering", ACM Trans. Inf. Syst., ACM, 2008 vol. 27, pp. 1-27

[23]. Joseph S. Kong and Behnam A. Rezaei and Nima Sarshar and Vwani P. Roychowdhury and P. Oscar Boykin, " Collaborative Spam Filtering Using E-Mail Networks", IEEE CS Press, August 2008, pp. 67-63

[24]. P. Oscar Boykin and Vwani P. Roychowdhury, "Leveraging social networks to fight spam", IEEE CS Press, April 2005 ,vol. 38, pp. 61 – 68

[25]. Paul-Alexandru Chirita and J'org Diederich and Wolfgang Nejdl, "MailRank: using ranking for spam detection", Proc. 14th ACM Intl. Conf. Information and knowledge management, ACM, 2005, pp. 373-380

[26]. Adam Berger and Vincent Della Pietra and Stephen A. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing", 1996

[27]. Ion Androutsopoulos and Georgios Paliouras and Vangelis Karkaletsis and Georgios Sakkis and Constantine D. Spyropoulos and Panagiotis Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach",??---vol: vcs.CL/0009009, 2000

[28]. Stop Comment Spam and Trackback Spam << Akismet http://www.akismet.com/ as on

25 April 2011

[29]. The Official CAPTCHA Site http://www.captcha.net/ as on 25 April 2011

[30]. Gordon V. Cormack and G'Jos omez Hidalgo'e Mar'ia and S'anz, Enrique Puertas, "Spam Filtering for Short Messages", Proc. 16th ACM Intl. Conf. on Information and Knowledge Management, ACM, 2007, pp. 313-320

[31]. Baoning Wu and Kumar Chellapilla, "Extracting Link Spam Using Biased Random Walks From Spam Seed Sets", Proc. 3rd AIRWeb Intl. workshop on Adversarial Information Retrieval on the Web, ACM, 2007, pp. 37-44

[32]. Bin Zhou and Jian Pei, "Link spam target detection using page farms", ACM Trans. Knowl. Discov. Data, ACM, 2009, vol. 3, pp. 1-38

[33]. Zoltan Gyongyi and Pavel Berkhin and Hector Garcia-Molina and Jan Pedersen, "Link Spam Detection Based on Mass Estimation", Proc. 32nd Intl. Conf. on Very Large Data Bases, VLDB Endowment, 2006, pp. 439-450

[34]. What The Trend. Find out WHY terms are trending on Twitter http://www.whatthetrend.com/ as on 25 April 2011

[35]. Wikipedia, the free encyclopedia http://www.wikipedia.org as on 25 April 2011

[36]. Google Search http://www.google.com as on 25 April 2011

[37]. http://www.dbpedia.org as on 25 April 2011

[38]. Home | OpenCalais http://www.opencalais.com as on 25 April 2011

[39]. Home | OpenAmplify . com http://www.openamplify.com as on 20 November 2010

[40]. Twitter Grader | Get Your Twitter Ranking http://twitter.grader.com as on 15 March 2010

[41]. Twitter API Wiki / FrontPage http://apiwiki.twitter.com/ as on 9 October 2011

[42]. Five Filters http://www.fivefilters.org/ as on 9 October 2010

[43]. Developer's Guide - Google Web Search API – Google Code http://code.google.com/apis/websearch/docs/ as on 5 January 2011

[44]. WordPress : Blog Tool and Publishing Platform http://www.wordpress.org/ as on 25 April 2011

[45]. What the Hashtag?! - the user-editable encyclopedia for Twitter hashtags http://www.wthashtag.com/ as on 23 October 2010

[46]. Chris Grier, Kurt Thomas, Vern Paxson, Michael Zhang "@spam: the underground on 140 characters or less", In Proceedings of the 17th ACM conference on Computer and Communications Security October 2010

[47]. Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng "Why We Twitter: Understanding Microblogging Usage and Communities" In Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007

[48]. Alexander Pak, Patrick Paroubek "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In Proceedings of the 7th conference on International Language Resources and Evaluation LREC'10 Valletta, Malta: European Language Resources Association ELRA, May 2010

[49]. Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas "Short text classification in Twitter to improve information filtering", In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval at UniMail, Geneva, Switzerland, July 2010