Improved Feature Selection for Better Classification in Twitter

Saumya Goyal
P.G. Student
JMIT Radaur
Yamunanagar

Shabnam Parveen
Assistant Professor
JMIT Radaur
Yamunanagar

ABSTRACT

Social networks are widely used as a communication tools by millions of people and their friends. In today's era everybody is online and use social network for interaction, to gain knowledge, for business purpose, politics and many more. But along with positive approach of using these tools some infect many negative approaches are also applied on these tools for executing malwares and spam messages. Spam on twitter has become one of the most trending topics of research in recent years. And many researchers have done work on it but make some very complex structure to detect spam but still cannot achieve that level of accuracy in detection. So to gain the greater level of accuracy and to reduce the complexity of structure this work proposes a simplified model to detect the spam tweets which are spread by unauthorised users or by spammers. And this is analysed by feature extraction and applying classifiers. The text and content attribute features are extracted by pre-processing and forming a feature vector matrix. Moreover K-nearest neighbour (KNN) and decision tree two classifier algorithms are applied to show the comparative results. The results are evaluated with False positive rate (FPR), F- measures, True positive rate (TPR) and accuracy with improved detection results.

GENRAL TERMS

K-nearest neighbour, Decision tree classifier algorithm, Pre-processing, Social network, Spam detection

1. INTRODUCTION

Internet has become an indispensable tool for communication, because of its fast speed and low cost. Often, search engines are the starting point for browsing on Internet. So the results of raking for a given query is highly important for commercial websites. Web spamming degrades the quality of the results retrieved by the web search engine.

[1]Due to the developed technology, an online social network turns into extremely important network to exchange and share information with each other. The number of online social networks have been entrenched and used by several different users. Twitter is one of the most popular social media as it permits user to mail and read innumerable posts associated to text. Today thousands and millions of users [2] share their information to others and there are approximately 400 million tweets every day [3].It is a common characteristic in social network that several individual characters [4] have particularly same influence on each other. Huge number of available users and quantity of exchanged data on the Twitter social networks heads to lot of cybercrime activities by spammers whose purpose is to extend spam messages through URLs of associated websites. [6] The motivation behind spam is to deliver information to the recipient that contains a payload like advertising for a (likely worthless, non-existent, or illegal) product, promotion of cause, bait for a fraud scheme, or computer malware designed to hijack the recipient's computer. As it is so cheap to send information, only a very small fraction of targeted recipients — possibly one in ten thousand or fewer-need to respond and receive to the payload for spam to be profitable to its sender[5].

The popularity of twitter is possibly due to its limitation. User posts are limited to 140 characters, and the privacy model is highly limited: a whole account is either private (only sharing posts with friends) or public, and most users select "public," sharing all of their content freely to world. Moreover following a user is not necessarily reciprocal: because all tweets are public, following a user merely subscribes a follower to their public tweets and thus users are encouraged to follow individuals, they do not know personally. This led to many celebrities using Twitter as a medium of connecting with their fans, because they can update their millions of fans with a single 140-character tweet.

Twitter has become a popular aspect in social networking spams [7] due to its susceptibility and vulnerability to attacks. This Twitter spam exploits the users with attracting messages such as, "Just saw this photo of you" which is followed by link that, takes user to an unauthorized site that uploads malware onto the user's computer [8]. In certain scenarios, by taking welfare of the phishing techniques [9]-[10], the messages may seem to come from one of the usual friends. Attackers or intruders uses Twitter to mediate coded update messages to users already infected by lawless code to handle botnets [11], which are groups of ruined computers that can be directed to notify various users who send spam or causes an attack over websites with polluted traffic.

In twitter social networking spam [6], spammers create fake accounts in order to steal the private data or to circulate commercial ads in social network for individual's benefit which affect the overall safety and performance of social networking. The main challenge is to identify twitter social spamming accounts created by spammer as their behaviour would have many varieties with much larger feature space which makes it difficult to detect. The important problem in detecting tweet social spammers is their dynamic nature, which makes it difficult. In traditional system, the performance is constant by applying direct conventional systems, as the spammers get develop new, more elusive techniques to avoid their detection.[13] Spam detection in social networks is relatively recent area of research. Most of the researches in this area follow the same general method of detection: 1) use empirical study to select some structural or textual features to examine; 2) use classification and machine learning techniques with these features to find patterns across users and messages;

3) evaluate whether models based on patterns are effective in detecting unwanted behaviour. Many researchers have introduced much complex structures or hybrid structures to remove this problem to achieve the accuracy in spam detection. [1][12] has introduced a hybrid model and extract numerous of features to detect the activity of spammers. Both clustering and classification techniques are used to detect spam and to enhance performance which make it more complex and time consuming system, along with this have not achieved that level of performance.

In order to solve these problems, this paper has proposed a simplified spam detection system. Initially, data is collected and text and content features are extracted by pre-processing. Then applying classifiers a performance matrix has evolved and deduce better performance of detecting spam and non-spam tweets.

2. PROPOSED WORK

In this model initially live tweets are collected through the Twitter API to perform the spam and non-spam detection and accordingly labelled the gathered data manually as no one provide the pre labelled data. We have created our own data sets which have current live data of twitter. The marked trending topics are used to detect both non-spam and spam tweet results. The main process that needs to be carried out is the extraction of a set of features from collected topics data. The text and content features of each and every twitter post is extracted and then pre-processed to form a feature vector matrix. Then final tweets which are characterized by a set of features is given to the K-nearest neighbour and decision tree classifier to train its model in detection of both spam and non-spam tweets. The whole architecture of the proposed system is shown in Fig. 1

A. Extractor/Reader

In order to correctly classify spam and non-spam user from trending tweets, then the tweets have to be named in an appropriate manner. Then the most important features of the tweets are extracted to improve the detection of illegal tweets and, are scarcely manipulable. The following features are important to detect spam and non-spam tweets results.

From every raw tweets, we have calculated the words in the tweets, profile name, hashtags, URLs, remaining words and count the number of words per tweet.

B. Pre-processor

Classifier algorithm cannot be directly interpreted the text. Documents should first be transformed into a representation state for the classification algorithms to be applied. In order to transform text into a feature vector, pre-processing is needed. This stage consists of identifying feature by feature extraction and feature weighting. The main goal of feature extraction is to transform a message from text format into a list of words as feature set. And this is done in four steps:

- Special Symbol Removal- in this all the special symbols such as !, :, &, * and many more will get removed as it has no effect on the sense of our statements.

- Stop Word Removal- stop-word removal consists in eliminating *stop-words*, i.e., words which provide less or no information to the text analysis. Common stop-words are articles, prepositions, conjunctions, pronouns, etc. Other stop-words are those having no statistical significance that is, those that typically appear very frequent in sentences of the considered language, or in the set of texts being analysed, and can therefore be considered as noise [14]. The authors in [15] have shown that the 10 most frequent words in texts and documents of the English language are about the 20–30% of the tokens in a given document. In the proposed system, the stop-word list for English language.

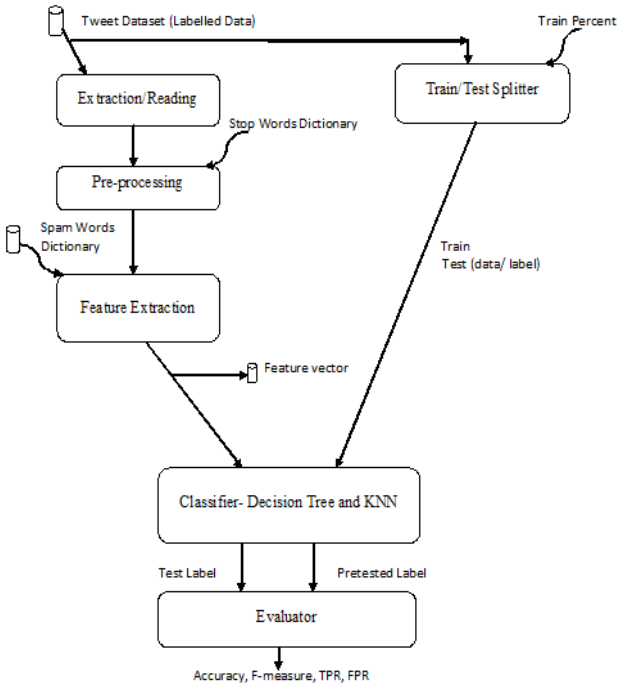

Fig1: Architecture of proposed system for tweet data

- Lower Case Conversion- in this it converts whole data into the lower case, by which it can do similarity match more easily and reduces the complexity. All the upper case letters get converted into lower case letters.
- Stemming- in this step all the affixes (prefixes and suffixes) are get removed from the words and converted them into its original or basic word like used, useful, using, uses etc into use.

After performing all these steps, whole dataset is converted into tokens and further get processed. Tokens contain all the basic words and required material which system requires for detecting spam and non-spam tweets.

C. Feature Extraction

In this a vector space is created which contain all the features which are required to be extracted for the spam detection. Basically text and content attribute features are extracted and spam word file is also loaded which contain all those words which are considered as spam in the corresponding dataset like abuse words, irrelevant URLs etc. It is dynamic in nature as we can expand it according to expansion of our dataset.

Following are the features which are extracted:

Table 1. Text and Content Features

Features	Description
numsw	Number of spam words per tweet
wordcount	Total number of words per tweet
numurlpw	Number of URLs per word
numurl	Number of URLs per tweet

RTcount	Number of retweets count
HTcount	Number of hashtags per tweet

The Text and content features contain the matter or message which people wants to spread widely.

For Example: RT @Martin1Williams Graeme Souness encouraged by recent developments explains #DeshatGardi why #Rangers are unique http://t.co/UrFTm05qv0 http://t.co/C shows naked photos http://t.co/P http://t.co/Px http://t.co/5 #Discount

- wordcount: this feature counts the total number of words per tweet. If any tweet contains just two or three words then that tweet will be of no use and its only act as spam for users as it will have no useful information.

  From above example
  words in tweet are {‘ RT’, ‘@Martin1Williams’, ‘Graeme’, ‘Souness’, ‘encouraged’, ‘recent’, ‘by’ , ‘developments’, ‘explains’, ‘#DeshatGardi’, ‘why’, ‘#Rangers’, ‘are’, ‘unique’, ‘http://t.co/UrFTm05qv0’, ‘http://t.co/C’, ‘shows’, ‘naked’, ‘photos’, ‘http://t.co/P’, ‘http://t.co/Px’, ‘http://t.co/5’, ‘#Discount’ }
  So, wordcount = 23

  Basically it works on the counting of the total words present in per tweet and each url is considered as one word.

- Numsw: this feature count the number of spam words per tweet. If any tweet contain any one of the spam word which we have listed then it is spam for all.

  From above example
  Spam words in tweet are {‘#DeshatGardi’, ‘naked’, ‘#Discount’}
  So, numsw= spam words/ wordcount

  Numsw basically works on the basis of frequency of occurrence of number of irrelevant words per tweet.

- Numurl: it extracts the total number of urls per tweet. If any tweet contains large number of links then it will be spam because that tweet have irrelevant links and just try to mislead users.

  From above example
  URLs in tweet are {‘http://t.co/UrFTm05qv0’, ‘http://t.co/C’, ‘http://t.co/P’, ‘http://t.co/Px’, ‘http://t.co/5’}
  Numurl = 5

  In this feature system observe that in all what is the frequency of URLs are there per tweet, if there are large frequency then it might be spam as it contains malicious urls which are only used to steel users information.

- Numurlpw: extract number of URLs per word, if there are more URLS per word then automatically there is something wrong in that link and will be spam for all.
  From above example
  URLs in tweet are {‘http://t.co/UrFTm05qv0’, ‘http://t.co/C’, ‘http://t.co/P’, ‘http://t.co/Px’, ‘http://t.co/5’}
  Numurl = 5
  So, numurlpw = numurl/ wordcount

  Numurlpw works on the average of occurrence of number of urls per word in each tweet.

- RTcount: it extracts the frequency of the retweet count done by any user.

  From above example
  On the basis of keyword RT and @ system observe whether following tweet is retweeted or not {‘RT’, ‘@Martin1Williams ‘}

- HTcount: it extracts the number of hashtag per tweet. Hashtag is started with the symbol # . If there are only hashtags in tweet then it is spam as contain no information, spammer only tweet this to follow these trends.

  From above example
  Hashtags in tweet are {‘#DeshatGardi’, ‘#Rangers’, ‘#Discount’}
  So, HTcount = no. of hashtags/ total words

  HTcount work on the average basis, that on an average per tweet how many hashtags are there.

A vector space model is a numerical representation used for analysis of text in various fields based on the probability measurement evaluation as represented in above table. In this a matrix is formed on the basis of frequency of these features occurred per tweet.

D. Train/Test Splitter
The manually labelled dataset is divided into two sets, one is train set on the basis of which we train our system and second is test set on the basis of which we observe the performance of our system that how exact the results are. We split the data in such a manner so that it will show the best results.

E. Classification
To classify the majority of dataset we engaged supervised machine learning algorithms. These algorithms are first trained on the labelled data to develop classification models that are then applied to unlabelled data to predict which tweets are in the spam class and which are in the non-spam class.

K-nearest neighbours classifier
The k-nearest neighbour (*K*-NN) classifier is considered as an example-based classifier that means that the training documents are used for comparison rather than an explicit category representation such as the category profiles used by other classifiers. There is no real training phase. When new document required to be categorized, the k most alike documents(neighbours) are detected and if a large enough proportion of them have been assigned to a certain category then new document is also assigned to this category, otherwise not . Additionally, finding the nearest neighbours can be quickened using traditional indexing methods. To decide whether a message is spam or non-spam, we look at the class of the messages that are the closest to it. The comparison among the vectors is a real time process. This is the idea of the k nearest neighbour algorithm:

Step1. Training
Store the training messages.
Step2. Testing
Given a message *x*, determine its *k* nearest neighbours among the messages in the training set. Whenever there are more spam's among these neighbours classify given message as spam. Otherwise classify it as not spam.

Decision Tree Classifier
Decision tree classifier is a method commonly used in data mining. The goal is to create a DT model and train the model so that it can divine the value of a target variable based on several input variables. In this every leaf represents a value of target variables given the values of the input variables represented by the path from root to the leaf.

A tree can be "trained" by splitting the source set into various subsets based on attribute value predefined. This process is repeated on each acquired subset in a recursive manner which is known as recursive partitioning. The recursion is accomplished when the subset at a node all has same value of the target variable or when splitting no longer adds value to the predictions.
In data mining, decision trees can also be described as the combination of computational and mathematical techniques to aid the illustration, categorization and generalization of a given allot of data. Data comes in records of the form:

(x , Y) = (x1, x2, x3, …., xk, Y)

The dependent variable, Y, is the target variable that we are trying to understand, generalize or classify. The vector x is composed of the input variables, x1, x2, x3…etc., that are used for task.

3. EXPERIMENTAL RESULTS

The performance of the proposed simplified classification approach for detecting the spam tweets, labelled collection of tweets are required which are pre characterised as spam and non-spam. It is observed that no such data collection is publicly available. So we labelled them manually by collecting them from web twitter application programming interface (API) method. The proposed approach is compared with the ELM based classifier approach only on the basis of content and text features [1]. Due to confidentiality policy of the Twitter, tweets cannot be reconstructed. This shows that there is no systematized collection with which to make tests and hence each researcher uses its own dataset. Therefore, every dataset has different characteristics, date and size. The standard parameters such as True positive rate (TPR), False positive rate (FPR), Accuracy and F-measures will be evaluated.

Table 2. Performance Comparison

Features	TPR	FPR	Accuracy	F-measure
ELM	94.2	2.7	94.3	93.89
KNN	97	1.8	97	98
DTree	100	0	100	100

A. True Positive Rate (TPR)

True Positive Rate is defined as the amount of the actual spam users who are correctly classified as spam users.

TPR is formulated as:

$$TPR = \frac{Tp}{Tp+Fn}$$

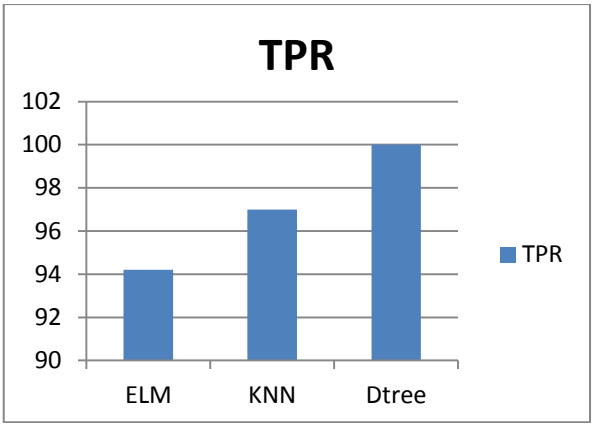

Fig: True Positive Rate for features

Above graph shows remarkable TPR evaluation has been carried out for features with classifiers presented above, which are implemented with twitter dataset collected by user. Results show that KNN and Decision Tree have remarkable improved TPR for twitter spam than ELM classifier.

B. False Positive Rate (FPR)

False Positive Rate is defined as the amount of the probability of wrongly classified results for spam and non-spam detection.

FPR is formulated as:

$$FPR = \frac{Fp}{Fp+Tn}$$

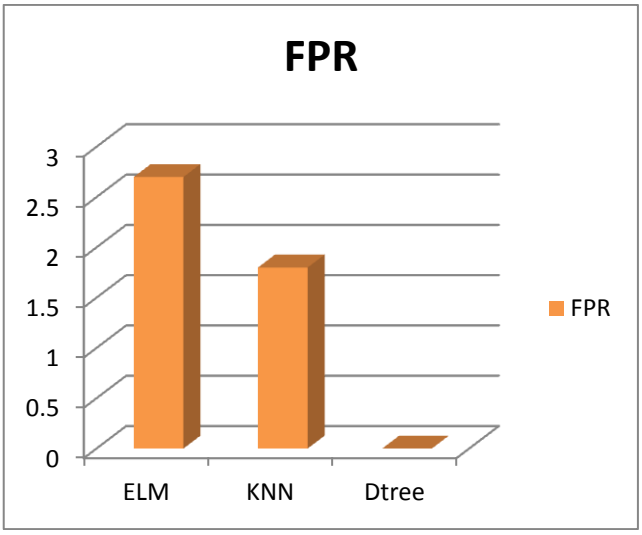

Fig: False Positive Rate for features

Extensive FPR evaluation of features for ELM and KNN, Decision Tree classifiers are presented in above graph. Results show that KNN and Decision Tree have less error rate for spam detection when compared with ELM classifier.

C. Accuracy
Accuracy is defined as overall correctness of spam detection results and is considered as the addition of the classification parameters (Tp + Tn) divided by the entire number of spam detection with classification parameters (Tp + Tn + Fp + Fn).

Accuracy is formulated as:

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn}$$

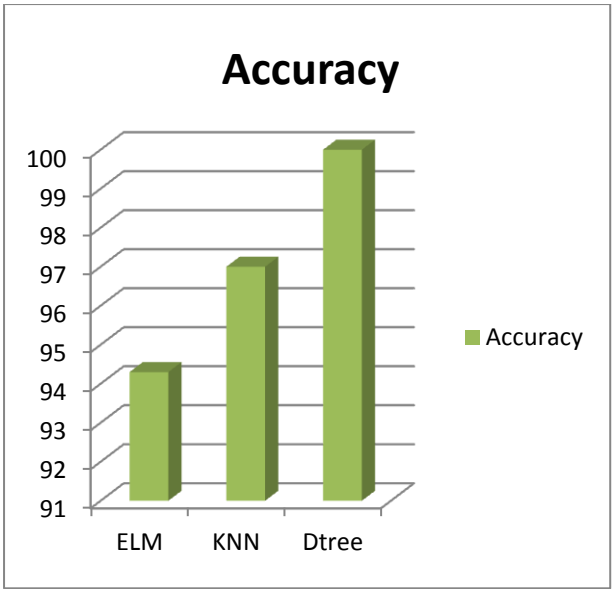

Fig: Classification Accuracy

An entire experimentation results for features with classifiers such as KNN, Decision Tree and ELM has been carried out as above. The results demonstrated that the KNN and Decision Tree classification method provides better spam detection rate with significant accuracy than ELM spam detection results.

D. F-measure

F-measure uses the result of both recall R and precision P

F-measure is formulated as:

$$\text{F-measure} = 2 * \frac{PR}{P+R}$$

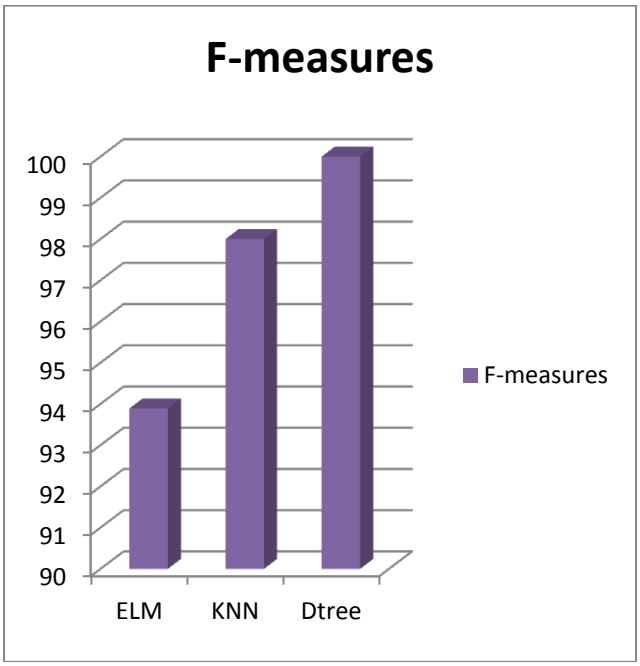

Fig: F-measures Accuracy

Above graph shows the F-measure comparison for standard measure of summarizing both recall R and precision P. The results shows that KNN and Decision Tree classification methods performs better spam detection rate than ELM spam detection results with higher F-measure value.

4. CONCLUSION AND FUTURE WORK

The main motive of this paper is to produce such a simplified system which reduces the complexity of the early proposed models and enhance the results of detecting spam and non-spam tweets. In this research work by extracting less but required features the performance of detecting spam tweets is achieved at much higher level and even on the basis of these feature extraction our model produces remarkable performance even at less percent of training set.

As we have performed it on limited dataset of tweets in future we can perform it on large dataset and not only of twitter but can be of any social site like Facebook, LinkedIn, pin interest and many more. As our model is dynamic in nature we can use it on any application without any modification. By just increasing the spam vocabulary and dataset we can perform it for any application.

This paper is mainly interested in detecting spam and non-spam tweets of twitter by extracting least number of features.

5. REFERNCES

[1] Saini Jacob Soman, Dr. S. Murugappan, "Detecting Malicious Tweets in Trending Topics using Clustering and Classification", 2014 International Conference on Recent Trends in Information Technology, IEEE

[2] Mashable. Twitter now has more than 200 million monthly active users. [Online]. Available:http://mashable.com/2012/12/18/twitter-200-million-active-users/

[3] Washington Post. Twitter turns 7: Users send over 400 million tweets per day. [Online]. Available: http://articles.washingtonpost.com/2013-03-21/ business/37889387_1_tweets-jack-dorsey-twitter.

[4] Il-Chul Moon,Dongwoo Kim, Yohan Jo and Alice O, "Analysis of twitter lists as a potential source for discovering latent characteristics of users," in CHI 2010 Workshop on Microblogging: What and How Can We Learn From It?, 2010.

[5] Neetu Sharma, GaganpreetKaur,et all, "Survey on Text Classification (Spam) Using Machine Learning", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5098-5102

[6] Gordon V. Cormack, *David R. Cheriton, "*Email Spam Filtering: A Systematic Review*",* Foundations and Trends ®in Information Retrieval Vol. 1, No. 4 (2006) 335–455©2008.

[7] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? Empirical evaluation and new design for fighting evolving twitter spammers", in 14th International Symposium, (RAID 2011), CA, USA, Proceedings in LNCS Series, Springer, Vol. 6961, pp. 318–337, 2011.

[8] N. Villeneuve, "Koobface: Inside a crimeware network", Munk School of Global Affairs, (JR04-2010), 2010.

[9] K. J. Nishanth, V. Ravi, N. Ankaiah, and I. Bose, "Soft computing based imputation and hybrid data and text mining: The case of predicting the severity of phishing alerts", in Expert Systems with Applications, Vol.39, Issue 12, pp.10583–10589, 2012.

[10] V. Ramanathan, and H. Wechsler, "phishGILLNET–phishing detection using probabilistic latent semantic

analysis", in EURASIP Journal on Information Security, 2012.

[11] J. Nazario, "Twitter-based botnet command channel", [Online].Available:http://asert.arbornetworks.com/2009/08/twitter-based-botnetcommandchannel

[12] Saini Jacob Soman, Dr. S. Murugappan, "Bayesian Probabilistic Tensor Factorization for Malicious Tweets in Trending Topics", 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE

[13] Grant Stafford, Louis Lei Yu, "An Evaluation of the Effect of Spam on Twitter Trending Topics", 978-0-7695-5137-1/13 © 2013 IEEE

[14] Y. Zhou and Z.-W. Cao, "Research on the construction and filter method of stop-word list in text preprocessing," in *Proc. 4th ICICTA*, Shenzhen, China, 2011, vol. 1, pp. 217–221.

[15] W. Francis and H. Kucera, "Frequency analysis of English usage: Lexicon and grammar," *J. English Linguistics*, vol. 18, no. 1, pp. 64–70, Apr. 1982.