Content Based Spam Classification in Twitter using MultiLayer Perceptron Learning

Arpna Dhingra
Department of Computer Science Engineering
Chandigarh University, Gharuan, Mohali (Punjab)

Shruti Mittal
Department of Computer Science Engineering
Chandigarh University, Gharuan, Mohali (Punjab)

Abstract - Social Networking Sites have created a boom in the industry of technology. Almost every second one out of three user is using one of the Social Networking Site. Due to high number of users, there comes a problem of Spam. Twitter is one of the vulnerable Social Networking Site and Twitter is facing a lot of problem due to malicious tweets sent by the spammers to lure the legitimate users. So, this paper represents a Content Based Spam Classification Approach i.e. MultiLayer Perceptron Learning for detection of Spam in Twitter. An analysis will be done using the MultiLayer Perceptron Learning and the results will then be compared with the already applied techniques for Spam Classification in Twitter.

Keywords- Content -Based, Spam Classification , Multilayer Perceptron Learning

I. INTRODUCTION

Social Networking Sites are becoming popular day by day. As much as Social Networking Sites ease the life of human beings, it also gives arise to various problems faced by users using these networking sites. There are various networking sites like Twitter, Facebook and LinkedIn etc. The major problem faced by users using these various networking sites is of Spam. Spam is any unwanted or prohibited behaviour that directly or indirectly violates the certain rules of any networking site. This paper mainly focuses on Spam Detection in Twitter using Multi-Layer Perceptron Learning. Till date, Twitter has faced

various problems due to spam which is created by various spammers to earn and fill their pocket. The main motive of any spammer is to lure the legitimate user towards his/her malicious spam. According to Twitter Policy, there are various tactics that are considered as Spam:

- Post harmful and malicious links
- Abusive replies to various users
- Posting duplicate and unrelated data for tweets
- Posting about current topics for seeking attention

Though, Twitter has followed some security measures to prevent spam but spammers are finding more and more techniques to trap legitimate users. So, basically this paper revolves around the technique of detecting spam in tweets. By fetching certain live tweets, it will classify the tweets in various content based features and then tell whether a tweet is a spam or not.
Multilayer Perceptron is a type of artificial neural network which has a certain set of data to be put as input onto a certain set of data to be put as output. There can be any number of layers in Multilayer Perceptron and each layer is connected to its successor. Nodes in Multilayer Perceptron are denoted as neurons. So, basically it is a type of finite acyclic directed graph.

The latter part of this paper described as: Section II throws some light on related research done this field. Section III explains the spam detection method and Section IV concludes the paper.

II. RELATED RESEARCH

Research in this field is going at a very fast pace. The authors in [1] have focused on an Integrated Approach in Spam Classification on Twitter using URL analysis, NLP and Machine learning Techniques. The combined approach has given better results and more accuracy rather applying all techniques alone.

Spam Detection on Twitter in [2] is done using traditional classifiers. The authors have discussed some user based as well as content based features and then have used them for spam detection. Random Forest Classifier has given the best results among SMO, Naive Bayes and K-NN neighbor.

The authors in [3] have detected Web Spam using Multi-Layer Perceptron Neural Network. MLP Neural Network provides flexibility and can easily accommodate web spam patterns. WEBSPAM-UK2006 and WEBSPAM-UK2007 datasets are used and various experiments are being performed.

The authors of [4] have performed Sentiment Analysis of Twitter Data. This paper has used lexicon-based as well as learning based techniques which are further used for Sentiment Analysis. Also, the authors have discussed various issues and challenges faced during the analysis of data.

The authors of [5] have followed a certain approach for detecting malicious tweets. Firstly, they have collected some data of twitter regarding trending topics and then labeled the tweets. Feature extraction is done and using FKM clustering is performed. Then the clusters are classified and malicious tweets are distinguished.

So, from the related research, it is concluded that various authors have used several techniques in alone as well as in collaborative manner but for detection of spam in tweets, Multilayer Perceptron is not used. So, we have used this technique for our research.
Also, authors have also proposed methods to detect malicious tweets using traditional classifiers as well by using clustering methods.
Authors have used Multilayer Perceptron in detection of web spam and not tweet spam. So, this reason got us inquisitive to do our research.

III. SPAM DETECTION METHOD

A. Features used for deciding spam

i. Number of Unique Hash-tags(#) : Hash-tags are basically used by spammers who intend to promote or sell some items online. A legitimate user may use the hashtag just once for promotion but a spammer will use the same hashtag again and again.

ii. Number of Unique URL's: Spammer uses the same URL again and again to trap a legitimate user. When the same URL is put in a tweet, the chances of getting it clicked by any legitimate user will be increased and spammer will earn all the benefit.

iii. Number of Unique @: Spammers may use same @ feature more than once. So, if a user has so many @ feature mentioned, there is a chance that he/she will be a spammer over the cover of legitimate user.

B. Proposed Methodology

The block diagram for proposed methodology is:

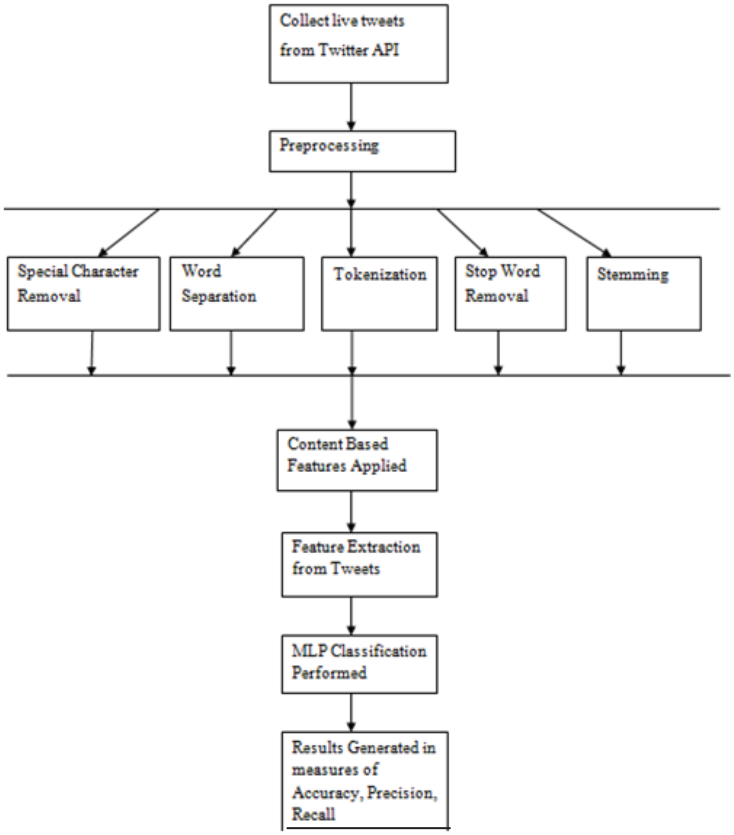

Fig.1. Block Diagram

i. Collection of live tweets through API's: The tweets are collected live through the API created. Every user can create its own API and each user get its own Consumer key or API key(c_key), Consumer secret or API secret(cs_key), Access Token and Access Token Secret(at_key). Then with the help of programming in python, tweets are displayed on screen. API's are generally set of routines, protocols etc. which are used for building a software application. If compared with accessing of databases or computer hardware, API eases the work of for any programming components. Various plugins are available for accessing API of twitter.

ii. Preprocessing: The data collected is then preprocessed. Basically, our system uses special character removal, word separation, tokenization, stop word removal and stemming. Special characters like <, = ,>, \\ etc. are removed. Also, big words are separated into small ones and then the whole data collected is converted into small tokens. So, the basic motive of preprocessing is to sieve the important data from the collected one.

iii. Feature Extraction: Our system mainly uses three content based features i.e. number of unique hashtags, number of unique URL's, number of unique @. So, by extracting features on these basis, all the tweets will get their own weights as well cost. Perceptron Learning has its own method to calculate weight as well as cost. In this research, the tweets already extracted will have certain weights as well as cost. The weights can be defined on the basis of content features like if a tweet having hashtags,

iv. Classification: Multilayer Perceptron learning is used to classify the data received after the whole process. We make our system learn in the language of neural network and then classify the tweet as a spam or a legitimate one. Algorithm for MLP used is:

i. Start

ii. Define words from the document for initialization
iii. Extract all words in the string which have length defined in a specified bounds
iv. Define probability of a document having features as well as label
v. Calculate probability of each feature given in a label
vi. Weight document probability by label probability doc_prob * label_prob
vii. Create list of all the probabilities
viii. Initialize list to store predicted class i.e. pred_class = [ ]
ix. Calculate distance with respect to training data distances.append((calc_dist(di,dj,dist), ij))
x. Define k-neighbors for class i.e. k_nn = sorted(distances)[:k]
xi. Calculate distance of every function used in data i.e. calc_dist(di,dj,i=1)
xii. Create an array to store the evaluation result
xiii. Increment the correct as well wrong prediction by 1

```
if x == 0:
eval_result[0] += 1
else:
eval_result[1] += 1
```

xiv. Set runtime for predictions i.e. start = time.clock()
xv. Run MLP Classifier for each k and distance function
xvi. Store the result and evaluate the predicted data
xvii. Assign evaluated result to classification result

```
pred_class = mlp(K[i], dtrain, dtest,   dtr_label, dist_fn[j])
eval_result = evaluate(pred_class-true_class
results.append(eval_result[0])
results.append(eval_result[1]results_mlp=[((results[0]+results[1])//4),(results[0]+results[1])-
((results[0]+results[1])//4)]
```

xviii. Print result on the screen
xix. Stop

Also for comparing the results , Naïve Bayes algorithm has been used on the same dataset and then accuracy, precision and recall is calculated. The values derived by MLP algorithm over Naïve Bayes Algorithm is better. The algorithm used for Naïve Bayes is defined below:

i. Start
ii. Define loadcsv(filename)
iii. Define splitDataset(dataset, split ratio)
iv. Define separateByClass(dataset)
v. Calculate mean of numbers and return ratio of sum of numbers to length in float of numbers
vi. Calculate standard deviation of numbers
vii. Define summarize dataset including both mean and standard deviation
viii. Define summarizeByClass including whole summarize dataset
ix. Define calculateProbability having variables x, mean, sd
x. Define calculateClassProbabilities
xi. Define predict with input vector
xii. Define getPredictions with summaries and test set
xiii. Check accuracy for defined test set
xiv. Print train and test set values for different split ratios defined
xv. Stop

IV. RESULTS

The results for tweets classified are defined in terms of accuracy, precision, recall. The values generated are presented in tabular form. So, for the above values of tweets and spam, simultaneous values of Precision, Recall and Accuracy has been calculated. Below are the tables and related bar graphs for precision, recall and accuracy are given below. The basic formulas to calculate precision, recall and accuracy are:

$$\text{Precision} = TP / TP + FP$$

$$\text{Recall} = TN / TN + FN$$

$$\text{Accuracy} = TP + FP / TP + FP + TN + FN$$

Table 1.1 Different Number of tweets taken and spam identified

Number of tweets taken	Total Test Records	Number of correctly identified Spam
90	88	66
75	63	48
100	38	29
120	18	14
20	9	3

Table 1.2 Values of Precision

True Positive	False Positive	Precision
40	26	0.606
20	28	0.416
15	14	0.517
2	1	0.666

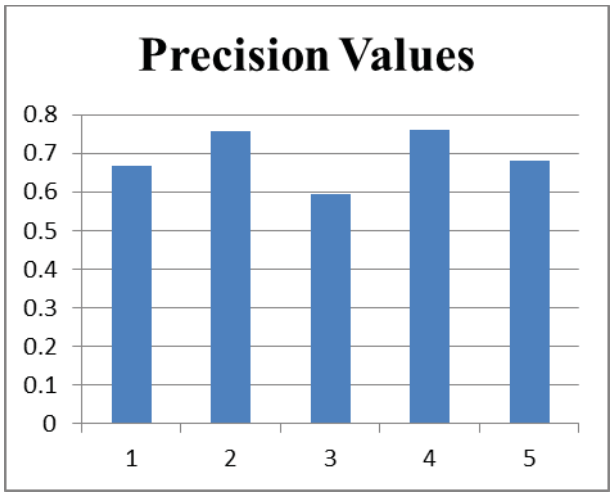

Fig.2. Precision Values

The precision values for MLP have turned out to be 75%. Similarly the values for recall and accuracy have been defined on the basis of TP, FP, TN, FN values. i.e. True Positive , True Negative, False Positive, False Negative.

Table 1.3 Values of Recall

True Negative	False Negative	Recall
20	4	0.833
10	17	0.370
60	11	0.845
90	16	0.849
10	7	0.588

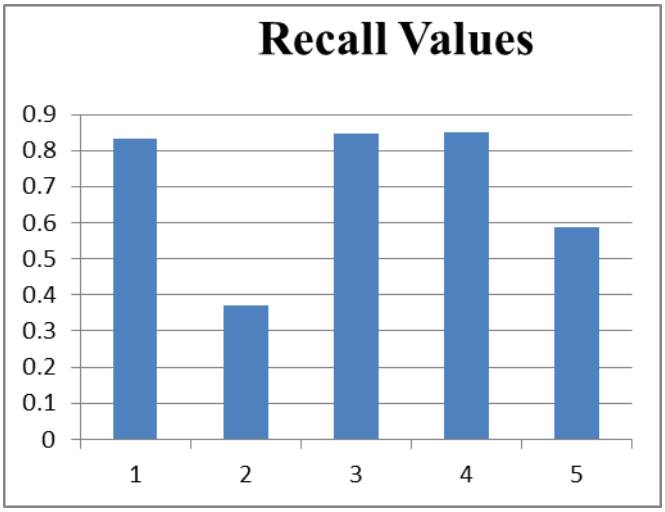

Fig.3. Recall Values

The recall value arrived for MLP technique used is 81%. These values are strictly based on hypothesis and may vary from system to system. Also, for the value of accuracy, bar graph has been plotted. The basic formulas to calculate precision, recall and accuracy are:

Table 1.4 Values of Accuracy

True Positive	True Negative	Accuracy
40	20	0.714
20	10	0.4
15	60	0.75
10	90	0.833
2	10	0.6

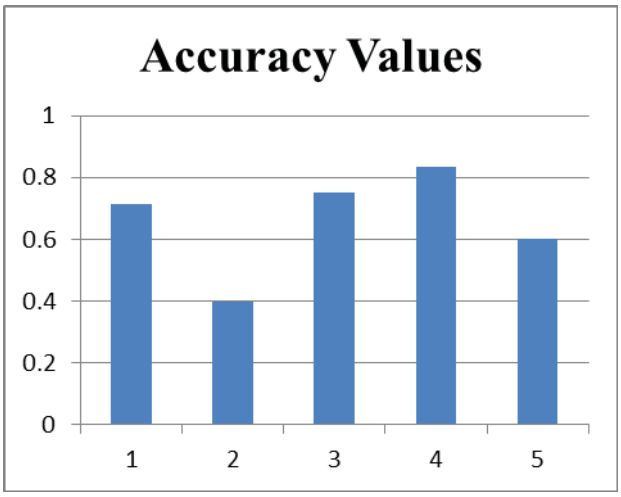

Fig.1.3 Accuracy Values

The accuracy achieved for MLP classification is 82%. Also, for comparing results, Naive Bayes is applied on the same dataset and results are then compared.

Table 1.5 Values of Split Ratio and Correctly Identified Spam

Split Ratio	Total Number in Test Data	Correctly Identified
0.4	88	71
0.5	88	55
0.7	88	33
0.8	88	20
0.9	88	8

True Positive	False Positive	Precision
6	3	0.666
25	8	0.757
35	24	0.593
60	12	0.759
30	14	0.681

Split Ratio is the ratio which explains how the data in dataset is divided or splitted into groups. This ratio has a value between 0 and 1.

Table 1.6 Values of Precision

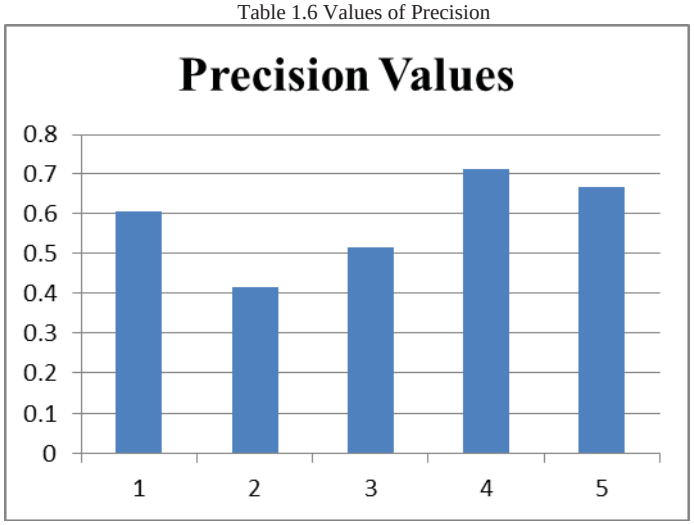

Fig.1.4 Precision Values

The precision value for Naive Bayes is 71% which is quite less than the precision value derived for MLP. Also, the other two values i.e. recall and accuracy have comparatively lesser values than the values derived with MLP classifier.

Table 1.7 Values of Recall

True Negative	False Negative	Recall
60	19	0.759
40	15	0.727
19	10	0.655
10	6	0.625
60	28	0.681

The bar graph plotted for recall in case of Naive Bayes system is:

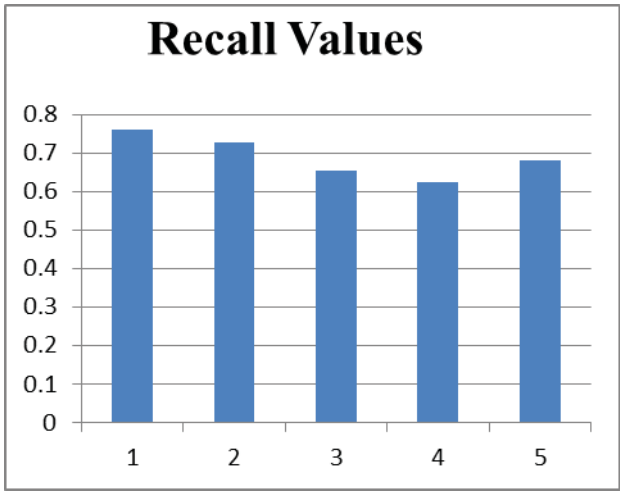

Fig. 1.5 Recall Values

The recall value arrived for Naive Bayes is 75% which is again lesser than MLP recall value. For accuracy, table generated is:

Table 1.8 Values of Accuracy

True Positive	True Negative	Accuracy
66	88	0.75
65	88	0.738
54	88	0.613
70	88	0.795
60	88	0.681

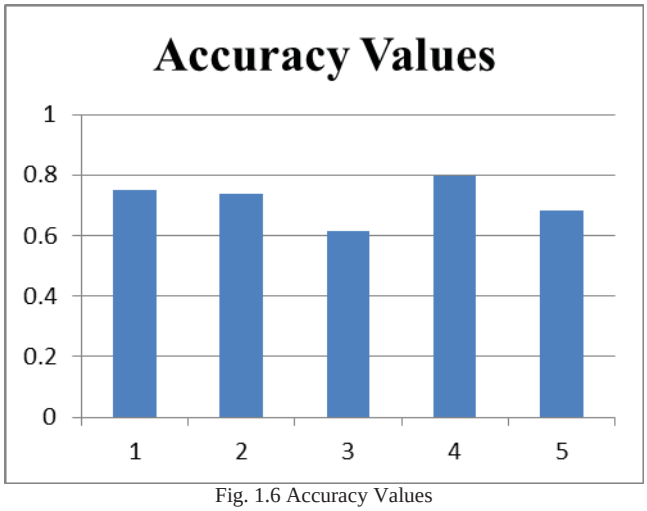

Fig. 1.6 Accuracy Values

The value derived for accuracy in case of Naive Bayes is 79% which is lesser than MLP.So, this proves that MLP has shown better results than Naive Bayes.

V. CONCLUSION AND FUTURE WORK

In this research, tweets has been classified into different categories of spam and fam. A defined methodology has been explained in previous chapters. This research first extracts the live tweets and then preprocess them to refine the tweets in a similar manner. Then, features are extracted and weights are stored in a file. After performing the MLP classification, accuracy, precision and recall values are calculated which explains that how accurately spam has been classified. This research work can be further extended by increasing the number of content features used. Also, this technique can be applied in some other context may be for any other social network. Also, comparison can be performed between various techniques and then results will be calculated.

REFERENCES

[1] Kamalanathan Kandasamy, P. K. (2014). An Integrated Approach to Spam Classification on Twitter using URL Analysis, Natural Language Processing and Machine Learning Techniques . *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science* (p. 5). IEEE.

[2] Cristina Radulescu, M. D. (2014). Identification of Spam Comments using Natural Language Processing Techniques. (p. 7). IEEE.

[3] M.McCord, M. (2011). Spam Detection on Twitter using Traditional Classifiers . (p. 7). Banff, Canada: IEEE.

[4] Sagar Bhuta, A. D. (2014). A Review of Techniques for Sentiment Analysis of Twitter Data. *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques(ICICT)* (p. 9). IEEE.

[5] Saini Jacob Soman, D. S. (2014). Detecting Malicious Tweets in Trending Topics using Clustering and Classification. *2014 International Conference on Recent Trends in Information Technology* (p. 6). IEEE.

[6] Kwang Leng Goh, A. K. (2013). MultiLayer Perceptrons Neural Network Based Web Spam Detection Application. (p. 5). IEEE.

[7] Radoslaw Michalski, P. K. (2012). Predicting Social Network Measures using Machine Learning Approach. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (p. 4). Asonam: IEEE.

[8] Kyumin Lee, B. D. (2012). Seven Months with the Devils: A Long- Term Study of Content Polluters. *IEEE* , 8.

[9] Ms.D.Karthika Renuka, D. M. (2011). Spam Classification based on Supervised Learning using Machine Learning Techniques. *IEEE conference on Machine Learning Techniques* (p. 7). IEEE.

[10] Neethu M S, R. R. (2013). Sentiment Analysis in Twitter using Machine Learning Techniques. *4th ICCCNT, 2013* (p. 5). Tiruchengode: IEEE.

[11] Kurt Thomas, C. G. (2011). Design and Evaluation of a Real-Time URL Spam Filtering Service. *2011 IEEE Symposium on Security and Privacy* (p. 16). IEEE.

[12] Asha S Manek, S. M. (2013). RePID-OK: Spam Detection using Repititive Pre-Processing. *2013 International Confernce on Cloud & Ubiquitous Computing & Emerging Technologies* (p. 6). IEEE.

[13] Krishna Chaitanya T, H. P. (2012). Analysis and Detection of Modern Spam Techniques on Social Networking Sites. *2012 Third International Conference on Services in Emerging Markets* (p. 6). IEEE.

[14] Kelton Costa, P. R. (2013). Comparison of the Techniques Decision Tree and MLP for Data Mining in Spams Detection to Computer Networks. *IEEE* , p. 5.

[15] Ze Li, H. S. (2011). SOAP: A Social Network Aided Personalized and Effective Spam Filter to Clean Your E- mail Box . *IEEE INFOCOM 2011* , 9.

[16] De Wang, S. B. (2013). Click Traffic Analysis of Short URL Spam on Twitter. *19th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing(CollaborateCom 2013)* (p. 10). IEEE.

[17] Ankita M. Ghate, L. M. (2015). Survey on Designing Framework for Analyzing Twitter Spammers using Forensic Method. *2015 International Conference on Pervasive Computing(ICPC)* (p. 4). IEEE.

[18] Amit A. Amleshwaram, N. R. (2013). CATS: Characterizing Automation of Twitter Spammers. *IEEE* , p. 10.

[19] Wang, A. H. (2011). Don't Follow Me: Spam Detection in Twitter. *IEEE* , 10.

[20] Ana C.E.S. Lima, L. N. (2013). Multi-Label Semi-Supervised Classification Applied to Personality Prediction in Tweets. *2013 BRICS Conference on Computational Intelligence & 11th Brazilian Conference on Computational Intelligence* (p. 9). IEEE.

[21] Hongyu Gao, Y. C. (2012). Towards Online Spam Filtering in Social Networks. *IEEE* , p. 16.

[22] Md. Saiful Islam, S. M. (2009). Modeling Spammer Behavior: Naive Bayes vs. Artificial Neural Networks. *2009 International Conference on Information and Multimedia Technology* (p. 4). IEEE.

[23] Sangho Lee, J. K. (May/June 2013). WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream. *IEEE Transactions on Dependable and Secure Computing Vol. 10* , 13.

[24] Grant Stafford, L. L. (2013). An Evaluation of the Effect of Spam on Twitter Trending Topics. *SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013* (p. 6). IEEE.

[25] Po-Ching Lin, P.-M. H. (2013). A Study of Effective Features for Detecting Long-Surviving Twitter Spam Accounts. *ICACT 2013* (p. 6). IEEE.

[26] Guangchen Ruan, Y. T. (2007). Intelligent Detection Approaches for Spam. *International Conference on Natural Computation(ICNC 2007)* (p. 5). IEEE.

[27] Chandra Shekar, S. W.-C. (2010). Mining Pharmaceutical Spam from Twitter. *2010 10th International Conference on Intelligent Systems Design and Applications* (p. 5). IEEE.

[28] Xiao mang Li, U. M. (n.d.). A Hierarchical Framework for Content-Based Image Spam Filtering. *IEEE* , p. 7.

[29] Claudia Meda, F. B. (2014). A Machine Learning Approach for Twitter Spammers Detection. *IEEE* , 6.

[30] Arushi Gupta, R. K. (2015). Improving Spam Detection in Online Social Networks. *IEEE* , 6.

[31] William Hua, Y. Z. (2013). Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter. *2013 Ninth International Conference on Semantics, Knowledge and Grids* (p. 8). IEEE.

[32] Beck, K. (2012). Analyzing Tweets to Identify Malicious Messages. *IEEE* , 5.

[33] Mohri Mehryar, Rostamizadeh Afshin, Talwalkar Ameet (2012). Foundations of Machine Learning : MIT Press
[34] Ganesan, Kavita. *A Brief Note On Stop Words For Text Mining And Retrieval*. 1st ed. Print.
[35] Hertzmann, Aaron, and David Fleet. *Machine Learning And Data Mining Lecture Notes*. 1st ed. Toronto: N.p., 2012. Print.