International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A Review on Techniques of Spam Classification in Twitter

Arpna Dhingra[1], Shruti Mittal[2], Harpreet Kaur[3], Arjinder Singh[4]

[1] **Arpna Dhingra**
Department of Computer Science & Engg.
Chandigarh University, Gharuan, Mohali(Chandigarh)
PG Student
arpna.appi24@gmail.com

[2] **Shruti Mittal**
Department of Computer Science & Engg.
Chandigarh University, Gharuan, Mohali(Chandigarh)
Assistant Professor
ershruti989@gmail.com

[3]**Harpreet Kaur**
Department of Computer Science & Engg.
Chandigarh University, Gharuan, Mohali(Chandigarh)
PG Student
harpysandhu91@gmail.com

[4]**Arjinder Singh**
Department of Computer Science & Engg.
Chandigarh University, Gharuan, Mohali(Chandigarh)
PG Student
arjinder8@gmail.com

Abstract: *The issue of spam in various Social Networking Sites has become very trivial these days. There are various users in the outside world who have turned into spammers in order to earn their benefits. So, to deal with the problem of spam, there arises techniques for Spam Classification through which a user can take preventive measures beforehand. This paper gives insight about various spam classification techniques like Naïve Bayes, Support Vector Machine, URL analysis etc. Also, this paper focuses on spam classification techniques used mainly for Twitter as Twitter is considered one of the vulnerable and sensitive Social Networking platform.*

Keywords: *Spam , Spam Classification , Naïve Bayes , Support Vector Machine , URL analysis.*

I. INTRODUCTION

Spam can be any undesirable and constrained behavior that may or may not breach the security policies of any network security. In other words, spam is any junk message or fake message sent by spammers to lure the legitimate users. There can be various intentions of spammers for sending these spam and unwanted messages like for advertising any product , intake of email addresses , accompany programs in partnership etc.
There are various kinds of Spam as described below:

- Email Spam : Email Spam is a type of spam sent through email. The links present in sent mails may mislead users to sites having phishing or malicious properties. The spammers play very smartly in case of email spamming. They assemble various email addresses from chat rooms, different websites ,blogs etc. and sell it to other spammers.

- Comment Spam : Comment Spam is a type of post in which spammer posts very abusive and offensive data in networking sites. In comment spam , spammer scatters the spam in the form of comments on blogs , forums , Wikipedia etc.

- Instant Messenger Spam : Instant Messenger Spam provides its users with a bulk of directory of its all users having analytical data like name , age , gender etc. Spammers looking for such susceptible data, gathers all details and by signing in, sends spontaneous and unsolicited messages which may contain any viruses , malicious links etc.

- Junk Fax : Junk fax is quite similar to email spam. The only difference lies is that spam in junk fax is received in the form of faxes through fax transmission. These are basically used for inflation in advertisements.

- Unsolicited Text Messages : Unsolicited Text Messages is a form of Mobile Spam. This spam type targets the message box of any user on his mobile phone but this type is less prevalent than email spam.

- Social Networking Spam : Social Networking Spam is a type in which spammed content is posted on social networks. Spam can be in the form of comments, tweets, chats, images etc. Basically, above explained types can be considered inside Social Networking Spam.

The latter part of this paper is as follows: Section II throws some light on related research done in this field. Section III explains the spam classification techniques and Section IV concludes the paper.

II. Literature Review

Research in this field is accelerating at higher bounds. The authors of [1] have served the spotlight 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.

For detecting spam comments by usage of Natural Language Processing Techniques, authors of [2] have used such procedures. They have designed an architecture for identification of spam comments.

Spam Detection on Twitter in [3] is done using traditional classifiers. The authors have conferred 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.

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

For detecting the malicious tweets, authors of [5] have followed a certain approach. 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. Clusters will be classified and malicious tweets are distinguished.

So, from the above literature review we conferred to a point that every technique has its own usage in different situations.

III.Spam Classification Techniques

3.1 ***URL Analysis*** **:**URL Analysis is basically performed in [1]. The authors have designed a system for spam classification. The URL from various tweets are first elicited and these are chiefly the short ones. These URL's are then converted to a longer form. For this the authors of [1] have used HttpURL Connection Class. All the URLs extracted are then compared with*:*

- Set of Blacklisted URL[1]
- Set of all already identified expressions[1]

The URL elicited are checked in the blacklist of URL. If one URL appears in the blacklist, that user will be defined as spammer and the whole process gets stopped as a legitimate user will never tweet any blacklisted URL.

In next stride, all URLs are checked for words defined as spam. The words taken as spam in [1] are bondage , showgirl, cybercore etc. The source of these spam words is www.urlblacklist.com [1]. If any expression contains such vulgar and obscene words, that URL will be a spam.

3.2 ***Natural Language Processing:*** Natural Language Processing is a technique in which machine learns the natural language and then performs every task which a human can do. In [1], authors have used two vital concepts for NLP:

- Remove Stop Words
- Stemming

For this, a set of spammed words have been exemplified. Stop words are eradicated first then stemming algorithm is implemented. If any word belongs to the list of spam words, that will be considered as spam.

Also authors in [2] have used a technique to identify spammed comments using NLP techniques. They have designed an architecture on the basis of which spam comments are classified. Firstly, comments having vulgar and obscene language, ambiguous comments are eradicated. Then, preprocessing of comments will be done. Thus, when any such property mentioned above is found the comments will be defined as spam or legitimate.

3.3 Naïve Bayes: Naïve Bayes is a machine learning classifier which uses Bayes theorem. The authors in [1] have used this technique for classification of spam. The set of labels is seized which encompasses both spam as well as legitimate data. So, on the basis of content based, authors in [3] have used traditional classifiers for spam detection. Algorithms like Random Forest, Naïve Bayes, Support Vector Machine, K- nearest neighbor have been taken into account and Random Forest have produced the best results [3].On the basis of content features like hashtags(#), @, Retweets etc. results are generated.

All these techniques have outperformed in one case or the other. We can't suggest one technique which is on top of everyone. Every method has it's own perfection as well as flaws.

IV. Conclusion

This paper has primarily focused on different spam classification techniques. The techniques like URL analysis, Naïve Bayes, Natural language Processing etc. have been annotated and every technique has its own process to classify spam. Also, this paper targets on spam classification in twitter as it is one of the vulnerable social networks.

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).*

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

[3] M. McCord, M. C. (2011). Spam Detection on Twitter Using Traditional Classifiers. ATC' 11, *Sept 2-4 ,2011 (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] SainiJacob 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] Ana C.E.S. Lima, L. N. (2013). Multi-Label Semi-Supervised Classication Applied to Personality Prediction in Tweets. *2013 BRICS Confernce on Computational Intelligence & 11th Brazilian Conference on Computational Intelligence* (p. 9). IEEE.

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

[8] Hongyu Gao, Y. C. (2011). *Towards Online Spam Filtering in Social Networks.* IEEE.

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

[10] Kyumin Lee, B. D. (2011). *Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter.* IEEE.

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

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

[13] 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). IEEE.

[14] Wang, A. H. (2012). *Don't Follow Me: Spam Detection in Twitter.* IEEE.

Author Profile

Arpna Dhingra received the B.E degree in Computer Science & Engg. From Punjab Technical University, Jalandhar and will receive the M.E. degree in Computer Science (with specialization Software Engineering) from Chandigarh University, Gharuan (Punjab). Her research interests include machine learning, UML diagrams.