International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 07 | July-2017 www.irjet.net p-ISSN: 2395-0072

Classification Methods for Spam Detection In Online Social Network

SUPRIYA RAMHARI MANWAR[1], Prof. P.D. LAMBHATE[2], Prof. J. S. PATIL[3]

[1] *ME Student, Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, Maharashtra, India*
[2] *Professor, Department of Computer and IT Engineering, Jayawantrao Sawant College of Engineering, Pune, Maharashtra, India*
[3] *Professor, Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, Maharashtra, India*

Abstract - *In the recent advanced society the online social networking sites like Twitter, Facebook, LinkedIn are very popular. Twitter, an online Social Networking site, is one of the most visited sites. Lot of users communicates with each other using Twitter. The rapidly growing social network Twitter has been infiltrated by large amount of spam. As Twitter spam is not similar to traditional spam, such as email and blog spam, conventional spam filtering methods are not appropriate and effective to detect it. Thus, many researchers have proposed schemes to detect spammers in Twitter, so need to identify spammers in twitter.*

Spam detection prototype system is proposed to identify suspicious users and tweets on Twitter. The proposed approach is to identify spam in Twitter using template, content, user based features to analyze behavior of user. Twitter API is used to get all details of twitter user and then generate the template. This template generated is then matched with predefined template. If suspicious behavior is analyzed, the account is considered as spam. However in case spam is not detected, the system collects 'content based' and 'user based' features from twitter account, by using the 'feature matching technique' to match features.

Algorithms used in the proposed system are supported by machine learning, which is used to match features and identify spam. Two Classification Algorithms, Naive Bayes and Support Vector Machine, are used for providing better accuracy and reducing execution time by the use of Template Matching. Public Dataset is collected from internet for providing training to Naive Bayes and Support Vector Machine classifiers.

Key Words: *Spam Detection, Twitter, Support Vector Machine, Naive bayes.*

1. INTRODUCTION

Recently the use of Social Network is increased tremendously to share people's views and ideas. Twitter is the social networking site used for sharing information about real world achievements. However nowadays we have observed that many people are using Twitter to do marketing and to spread spam massages in OSN (Online social Network).

Spammers have various kinds of motivations to spam the messages. For some people, the motivation can be financial gain; which is very clear from the tweets related to advertising a product or tweets by an online merchant to link to his website. Many a times these sellers may not be meticulous, and so they are prepared to disturb users by blocking their Twitter feed. Another kind of common type of spam is the tweets containing pornographic material or information of pornographic websites. In such scenarios our spam detection task could be viewed as a content filtering task.

Twitter does not allow pornographic material in profile, header or background images, but many accounts ignore this rule. This disregard for the Terms of Service could arguably be reason enough to find and remove such accounts. Whilst such content is viewed as lawful by some and some want to see it, it is many times a fascia for malware; links contained may be unsafe, with the risk of user's computer being infected with viruses.

The proposed novel approach is to detect spam in OSN. I have used Machine Learning Approach to classify given account and recognize the spammers. In Machine learning we need trained machines to predict the respective result to show spammers. Machine learning is divided into two parts:

Supervised Learning and Non-supervised Learning.

In Supervised Learning, we need to train the classifier. In Unsupervised Learning, we do not need to educate the classifier. However Supervised Learning gives better accuracy as compared to Unsupervised Learning.

In this paper, a description of twitter is given to identify the spam. In section II, literature survey of spam detection is done. Section III shows the proposed framework design. In section IV detailed description of classification process is described. Section V describes the dataset and predicted results. In section VI Graph are shown showing the

benefits and accuracy of the proposed solution. Finally, section VII gives the conclusion of the paper.

2. REVIEW OF LITERATURE

Paper [1], Spam is not as diverse as It Seems: Throttling OSN Spam with Templates Underneath, states that in online social network, spam is originated from our friends and thus it reduces the joy of communication. Normally spam is detected in text format. The system collects large amounts of data from online social network and that data is used for identifying spam. This identified spam is used for generating template. Whenever new stream of messages comes for identifying spam or not spam, those generated templates are used for matching with stream of messages, so it reduces execution time of identifying spam. That implemented framework is called as Tangram.

Paper [2], Detecting and Characterizing Social Spam Campaigns mentions that several online social networks are detected in internet. For identifying spam in online social networks, existing method uses the Facebook wall post. Crawlers are used for collecting wall post in particular Facebook user. Then this wall post filters and finally collects wall post which contains the URLs. This method differentiates wall post text and link which is mentioned in the wall. This method collects group from similar texture content and posts it including the same destination URLs. Post Similarity graph clustering algorithm is used to identify similarity between post and URL. Based on this malicious user and post is identified.

Paper [3], WARNINGBIRD: Detecting Suspicious URLs in Twitter Stream details about three modules, Data Collection, Feature Extraction and Classification. Under Data Collection, system collects tweets with URL by using Twitter Streaming API which is publicly available for getting data from twitter. In Feature Extraction, features are extracted from existing data. URL redirects chain length like feature collect system because attackers use long URL redirect chain to make analysis difficult. Suspicious URL on twitter is classified Based on the feature.

Paper [4], Suspended Accounts in Retrospect: An Analysis of Twitter Spam states that spam users continuously send abuse data in online social network. In this study, system first of all collects the 1.8 billion account data which is spam and analyzes web services like URL which contain abuse data. Based on the collected data we identify given account is spam or not spam.

Paper [5], Detecting spammers on social networks mentions that system collects user information like tweets, number of followers etc. This is done using Weibo API which is used for crawling. Feature extraction module uses two important features, Content based and User based features. In Content based feature, system identify number of posts and number of repost per day. User based feature extracts tweet post date, average number of messages and URL posted per day. Based on this feature SVM classifies instance. This binary classifier predicts whether user is spam or not spam.

Paper [6], Towards online spam filtering in social networks mentions that Online Social Networks (OSNs) are very much popular among Internet users. In case it is handled by wrong people, they are also effective tools for spreading spam campaigns. In this paper author present an online spam filtering system that can be used real time to inspect messages generated by users. The system can be deployed as a component of the OSN platform. Author proposes to rearrange spam messages into campaigns for classification instead of examining them individually. Although campaign identification is used for offline spam analysis, author applies this technique to support the online spam detection problem with sufficiently low expenses. Accordingly, this system adopts a set of fresh features that effectively distinguish spam campaigns. It drops messages classified as “spam” before they reach the recipients, thus protecting them from various kinds of fraud. The system is evaluated using 187 million wall posts collected from Facebook and 17 million tweets collected from Twitter.

3. EXISTING SYSTEM

Existing system used user-based and content-based features that are different between spammers and legitimate users. Then, they use these features to facilitate spam detection. Using the API methods provided by Twitter, they crawled active Twitter users, their followers/following information and their most recent 100 tweets. Then, we analyzed the collected dataset and evaluated our detection scheme based on the suggested user and content-based features. They show result by use of classifiers.

In Existing System required more execution time for identify spam in Twitter Data and that methods provide the less Accuracy.

Disadvantages of Existing System

1. It required more computational time for running classifier because while running they match training and testing instances.

2. System degrades the accuracy because system uses the classification only.

3. This application used in real time spam detection so it must have to provide better performance.

4. In classification, classifier identify spam based on training data. This approach not ability to identify new type spam.

4. PROPOSED SYSTEM

Detailed description of the system is discussed in this section.

System Overview

The aim of the proposed spam detection system is to detect the spam in Twitter by providing proper identification of spam in real time Twitter data. It provides accurate and the fast spam detection. In Existing System required more execution time for identify spam in Twitter Data and that methods provide the less Accuracy.

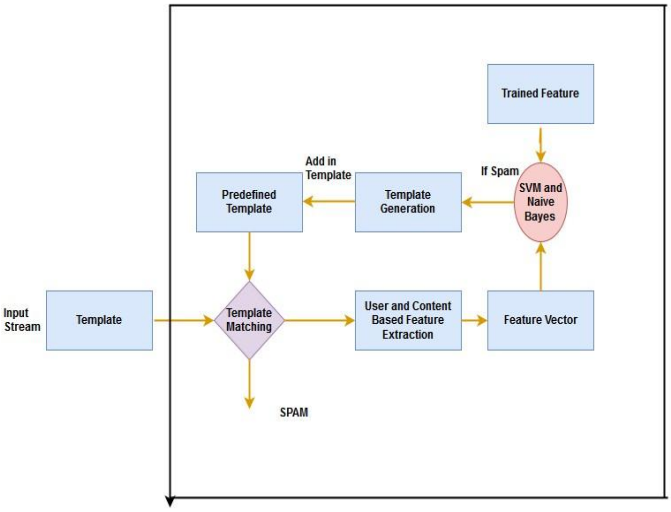


Figure 1. System Architecture

Module Description:

1) *Data Collection*: To fetch a data from twitter we need access of twitter, access obtained by creating a twitter Application. Whenever we create application we get four access keys from twitter.

They are four required for integrate twitter:

- Consumer Key
- Consumer Secret
- OAuth Access Token
- OAuth Access Token Secret

By using this key in Java program we are able to collect user data. System collects the input as twitter data and later use for template matching or classification using SVM.

2) *Template Matching*: Template contains bag of words. Given template matched with predefined template and identify spam or not spam user. If not spam then later we use twitter data for classification. If spam then given user is considered as a spam.

3) *Preprocess*: The twitter contain noise. That will decrease accuracy of the system so we need to remove noise from the twitter data.

4) *Feature Extraction*: We collect user based feature and content based feature from twitter data. User based feature contains user name, profile image, account details etc. Content based feature contains user tweets retweet etc. Based on this feature we train and test the model and identify spam using support vector machine and Naive Bayes algorithm.

5) *Classification*: SVM classification is essentially a binary (two-class) classification technique, which has to be modified to handle the multiclass tasks in real world situations. SVM and Naive Bayes classification uses features of twitter data to classify. This classification is uses trained twitter feature and classify testing twitter feature and identify spam or not.

6) Template Generation: If Support Vector Machine and Naive Bayes detected as spam then we generate template and given template added into predefined template

5. CLASSIFICATION

Input: A Twitter Feature
Dataset:
We used public SMS Spam Collection dataset which is available on internet. Dataset contains sentence with class label. We train Naïve Bayes algorithm and assigned label like ham and spam. Ham class label contains 4825 instances and spam class label contains 747 instances based on this instance, system predicts the given tweet is spam or not spam.
In stop word removal technique system uses the mallet LDA Stop word dataset. Mallet LDA contains list of stop words and that stop word compare with tweet and remove words which is present in dataset.

Output: class label (spam or not spam)

Process of SVM:-

1) Compute Score of input vector:
2) Kernel function (Radical basis function):
3) Class y = -1 when output of scoring function is negative.
4) Class y = 1 when output of scoring function is positive.
Parameter Xi ith value of input vector Yi ith value of class label a_i is the coefficient associated with the i^{th} training dataset b -Scalar value.

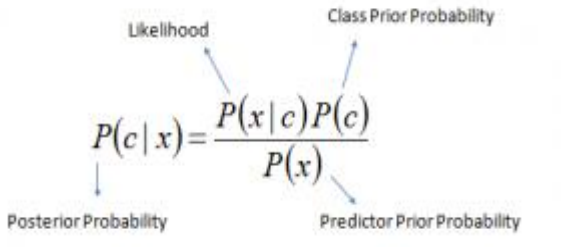
Process of Naive Bayes Therom: -

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

Algorithm for updated Naivebayes :

$$\tilde{p}(x,y) \equiv \frac{1}{N} \times \text{number of times that } (x,y) \text{ occurs in the sample}$$

- x_i includes the contextual information of the document (the sparse array) and y_i its class.
- N is the size of the training dataset.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

- P(c—x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x—c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor

6. RESULT AND ANALYSIS

We collect manually data using Twitter API and those data

used for feature selection and analyzing user account is spam or not spam.

Twitter Spam Percentage Graph

We perform spam detection on Facebook's twitter account and then fetch the tweets in Facebook account. Template matching to detect tweet spam or not spam. Then calculate percentage of spams by using given formula.

Percentage of spams=total no. of spam count / total no of tweet * 100

Result Table:-

Twitter Account	Spam Count	Not Spam Count	Total Count
Facebook	459.0	1641.0	2100.0
Gmail	252.0 11]	1848.0	2100.0
Linked In [15]	232	1596.0	2100.0

This table shows the output of spam detection, we analyze three Twitter account like Facebook, Gmail and LinkedIn. Gmail and LinkedIn accounts have less spam percentage as compare to Facebook twitter account. If spam percentage is less then that account is not spam.

Figure 2 shows the Facebook page in twitter how many spam tweets identified. Red color shows the spam tweet percentage and blue color shows the not spam tweet percentage. We collect tweet from twitter and remove the stop words from tweet and then apply naïve Bayes classification.

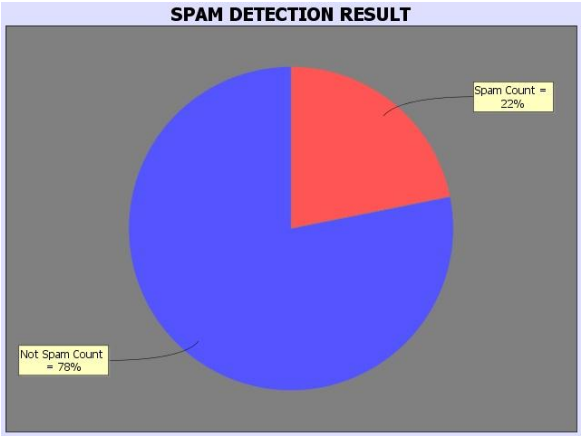


Figure 2. Twitter Spam Percentage Spam

Accuracy Graph

Figure 3 shows the accuracy comparison with SVM and updated naïve bayes. In previous system standard naïve bayes gives 93.7 but we use combination of entropy and naïve bayes which gives 97.4910394265233 accuracy. SVM is not giving better accuracy.

For analyze accuracy we used Weka tool. Naive bayes give 97% accuracy on spam identification and Svm give 56% accuracy.

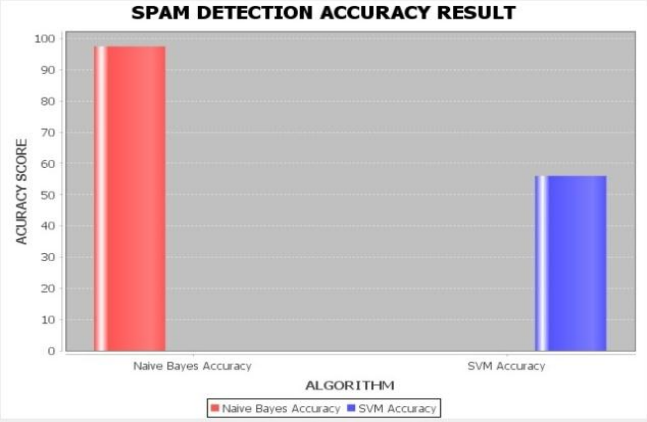


Figure 3. Spam Detection Accuracy Result

Execution Time Graph

Figure 4 shows the execution time required for Tweet collection, Stop word removal and classification. Tweet collection required more time as compare to others because it collects tweet from online twitter account and speed totally depend on internet speed.

Stop word removal technique remove the stop words from tweet and System compare tweet word with predefined stop word dataset.

In Implementation we are use java language so we get execution time in nanosecond by using System class. We convert given nanosecond into second and plot the graph.

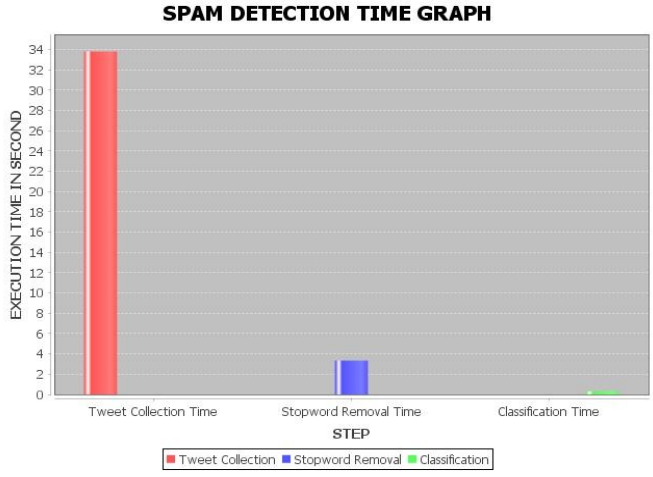


Figure 4. Spam Detection Time Graph

7. CONCLUSIONS

In this project we used template matching approach for identify given tweet is spam or not. There are two main factor of that project which is accuracy and execution time. For providing more accuracy we are using updated naïve bayes with the help of entropy and naïve bayes. For providing less execution time we are store trained data in main memory as well as choosing naïve bayes algorithm. The updated naïve bayes performs less process so that will reduce the processing time and improving performance of the system.

8. FUTURE SCOPE10

In future we are detecting spam on other online social networks like Facebook, Google+ and Linkedin etc as well as we also detects collusion and Sybil attack in twitter accounts. After that we give permission from twitter to remove spam accounts and tweets from twitter or other online social network.

9. ACKNOWLEDGEMENT

It is with the profound sense of gratitude that I acknowledge the constant help and encouragement from our guide Prof. P. D. Lambhate mam and Co-guide Prof. J. S. Patil mam and HOD Sir, Computer Engineering Department, also PG co-ordinator Sir and Principal Sir JSCOE, Hadapsar, for their sterling efforts, amenable assistance and inspiration in all my work. They have given in depth knowledge and enlightened me on this work.

10. REFERENCES

[1] H. Gao et al., Spam aint as diverse as it seems: Throttling OSN spam with templates underneath, in Proc. ACSAC, 2014, pp. 7685. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao.

[2] HongyuGao NorthwesternUniversity Evanston,IL,USA hygao@u.northwestern.edu, JunHu HuazhongUniv. ofSci. &Tech andNorthwesternUniversity junehu1210@gmail.com Detecting and Characterizing Social Spam Campaigns.

[3] S. Lee and J. Kim, WarningBird: Detecting suspicious URLs in Twitter stream, in Proc. NDSS, 2012, pp. 113.

[4] K. Thomas, C. Grier, V. Paxson, and D. Song, Suspended accounts in retrospect: An analysis of Twitter spam, in Proc. IMC, 2011, pp. 243258.

[5] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Yuanlong Yu, ChunmingRong,Detecting spammers on social Networks, Neuro-computing, http://dx.doi.org/10.1016/j.neucom.2015.02.047

[6]Hongyu GaoNorthwestern University,Evanston, ILUSA,hygao@u.northwestern.edu,Yan Chen Northwestern University Evanston, IL, USA ychen@northwestern.edu, Towards Online Spam Filtering in Social Networks.

[7] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, Design and evaluation of a real-time URL spam

filtering service, in Proc. IEEE Symp. SP, May 2011, pp. 447462.

[8] J. Mottl, Twitter acknowledges 23 million active users are actually bots, Tech Times, Aug. 2014 [Online]. Available: http://tinyurl.com/l755bvm.

[9] C. Kreibich et al., Spamcraft: An inside look at spam campaign orches- tration, in Proc. LEET, 2009, p. 4.

[10] C. Kreibich et al., On the spam campaign trail, in Proc. LEET, vol. 8. 2008, pp. 19.

[11] A. Pitsillidis et al., Botnet judo: Fighting spam with itself, in Proc. NDSS, Mar. 2010, pp. 119.

[12] Q. Zhang, D. Y. Wang, and G. M. Voelker,DSpin: Detecting automati- cally spun content on the Web, in Proc. NDSS, 2014, pp. 116.

[13] A. Ramachandran, N. Feamster, and S. Vempala, Filtering spam with behavioral blacklisting, in Proceedings of the 14th ACM Conference on Computer and Communications Security,2007.

[14] G. Stringhini, C. Kruegel, and G. Vigna, Detecting Spammers on Social Networks, in Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2010.

[15] C. Yang, R. Harkreader, and G. Gu, Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers, in Proc. RAID, 2011, pp. 318337.

[16] G. Stringhini, C. Kruegel, and G. Vigna, Detecting spammers on social networks, in Proc. ACSAC, 2010, pp. 19.

[17]http://www.cs.bu.edu/fac/gkollios/ada05/LectNotes/lect25-05.pdf

[18] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, Detecting spammers on Twitter, in Proc. CEAS, vol. 6. 2010, p. 12

BIOGRAPHY

Ms. Supriya Ramhari Manwar is currently pursuing M.E. Computer From Jayawantrao Sawant College of Enginnering, Savitribai Phule Pune University Pune, Maharashtra, India. She received her B.E. Computer Degree from Amaravati University. Her area of interest is Data Mining.