Identifying Tweet is Spam or Ham using Machine Learning Techniques

G. Nagarajan[1], C. Havinash [2], Y. Manoj Kumar[3], K. Suja Rajeswari[4]

[1, 2, 3] Student, Computer Science and Engineering, Panimalar Institute of Technology, Chennai,India.
[4] Assistant Professor, Computer Science and Engineering, Panimalar Institute of Technology, Chennai,India.

---***---

Abstract - *In the recent advanced society the online social networking sites like Twitter, Facebook, and LinkedIn are very fashionable. Twitter, a web Social Networking site, is one among the foremost visited sites. Lot of users communicates with one another using Twitter. The rapidly growing social network Twitter has been infiltrated by great deal of spam. As Twitter spam isn't almost like traditional spam, like email and blog spam, conventional spam filtering methods aren't appropriate and effective to detect it. Thus, many researchers have proposed schemes to detect spammers in Twitter, so got to identify spammers in twitter. Spam detection prototype system is proposed to spot suspicious users and tweets on Twitter. The proposed approach is to spot spam in Twitter using template, content, user based features to research behavior of user. Twitter API is employed to urge all details of twitter user then generate the template. This template generated is then matched with predefined template. If suspicious behavior is analyzed, the account is taken into account as spam. However just in case spam isn't detected, the system collects 'content based' and 'user based' features from twitter account, by using the 'feature matching technique' to match features. Algorithms utilized in the proposed system are supported by machine learning, which is employed 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 utilization 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 utilization of Social Network is increased tremendously to share people's views and concepts. Twitter is that the social networking site used for sharing about world achievements. However nowadays we've observed that a lot of people are using Twitter to try to Marketing and to spread spam massages in OSN (Online social Network).

Spammers have various sorts of motivations to spam the messages. for a few people, the motivation are often financial gain; which is extremely clear from the tweets associated with advertising a product or tweets by a web merchant to link to his website. Many times these sellers might not be meticulous and so they are prepared to disturb users by blocking their Twitter feed. Another kind Of common sort of spam is that the tweets containing pornographic material or information of pornographic websites. In such scenarios our spam detection task might be viewed as a content filtering task.

Twitter doesn't 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 seek out and take away such accounts. Whilst such content is viewed as lawful by some and a few want to ascertain it, it is repeatedly a fascia for malware; links contained could also be unsafe, with the danger of user's computer being infected with viruses.

In Machine learning we'd like trained machines to predict the respective result to point out spammers. Machine learning is divided into two parts:

1.1 Supervised Learning and Non-supervised Learning

In Supervised Learning, we'd like to coach the classifier. In Unsupervised Learning, we don't have to educate the classifier. However Supervised Learning gives better accuracy as compared to Unsupervised Learning.

In this paper, an outline of twitter is given to spot the spam. In section II, literature survey of spam detection is completed. 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 the Literature survey of spam detection is completed. Section III shows the proposed Benefits and accuracy of the proposed solution. Finally, section VII gives the conclusion of the paper.

2. REVIEW OF LITERATURE

[1] Spam isn't as diverse because 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 enjoyment of communication. Normally spam is detected in text format. The system collects large amounts of knowledge from online social network which data is employed for identifying spam. This identified spam is employed 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 named as Tangram.

[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 especially Facebook user. Then this wall post filters and eventually collects wall post which contains the URLs. This method differentiates wall post text and link which is mentioned within the wall. This method collects group from similar texture content and posts it including an equivalent destination URLs. Post Similarity graph clustering algorithm is employed to spot similarity between post and URL. Supported this malicious user and post is identified.

[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 form analysis difficult. Suspicious URL on twitter is assessed supported the feature.

[4] Suspended Accounts in Retrospect: An Analysis of Twitter Spam states that spam users continuously send abuse data in online social network. During 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.

[5] Detecting spammers on social networks mentions that system collects user information like tweets, number of followers etc. this is often done using Weibo API which is employed for crawling. Feature 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 postdate, average number of messages and URL posted per day. Supported this feature SVM classifies instance. This binary classifier predicts whether user is spam or not spam.

[6] Towards online spam filtering in social networks mentions that Online Social Networks (OSNs) are considerably popular among Internet users. Just in case it's handled by wrong people, they're also effective tools for spreading spam campaigns. During this paper author present a web spam filtering system which will be used real time to examine messages generated by users. The system is often deployed as a component of the OSN platform. Author proposes to rearrange spam messages into campaigns for classification rather than examining them individually. Although campaign identification is employed for offline spam analysis, author applies this system to support the web spam detection problem with sufficiently low expenses. Accordingly, this technique adopts a group of fresh features that effectively distinguish spam campaigns. It drops messages classified as "spam" before they reach the recipients, thus protecting them from various sorts 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 bonfire 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 up-to-date 100 tweets. Then, we analyzed the collected dataset and evaluated our detection scheme supported 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 which methods provide the less Accuracy.

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 supported training data. This approach not ability to spot new type spam.

4. PROPOSED SYSTEM

Detailed description of the system is discussed during this section.

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 therefore the fast spam detection. In Existing System required more execution time for identify spam in Twitter Data which methods provide the less Accuracy.

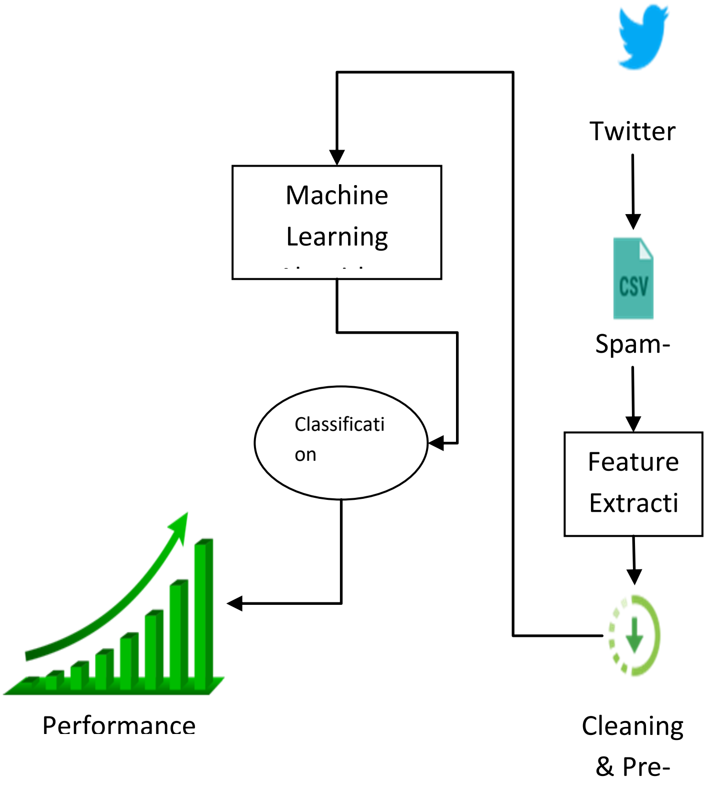


Fig - 1: System Architecture

In Fig 1 says using the csv data set we can preprocess the data , using the feature extration techniques the data can be cleaned and using the machine learning algorithm the data set will be trained and predict the accuracy higher.

4.1 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.

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.

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

3) 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.

4) 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.

5) 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 the value of input vector Yi it value of class label.

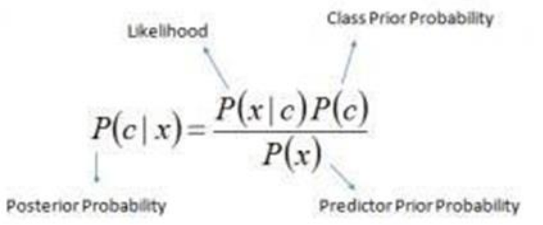
Process of Naive Bayes Theron: -
Bayes theorem provides a way of calculating posterior probability P (c|x) from P(c), P(x) and P (x|c). Look at the equation below:

Algorithm for updated Naive ayes :

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

- Xi includes the contextual information of the document (the sparse array) and yi 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 poterior 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

Table -1: Twitter Spam count and Non-Spam Count

Account	Spam Count	Non-Spam Count	Total Count
Facebook	459.0	1641.0	2100.0
Gmail	252.0	1848.0	2100.0
Linked In	232.0	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 than that account is not spam.

Fig 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.

Fig-2: Twitter Spam Percentage Spam

5.1 Accuracy Graph

Fig 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 97.4910 accuracy. SVM is not giving better accuracy.

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

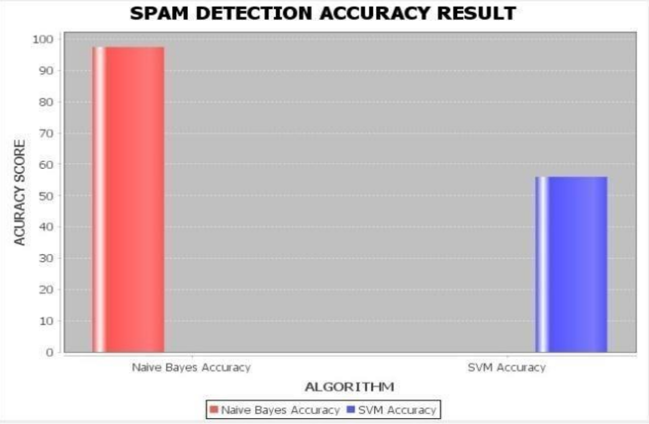


Fig-3: Spam Detection Accuracy Result

5.2 Execution Time Graph

Fig 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.

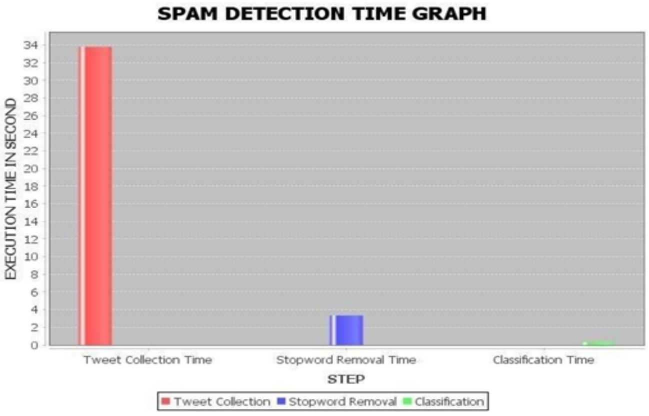


Fig- 4: Spam Detection Time Graph

7. CONCLUSIONS

In this paper, template matching approach for identify given tweet is spam or not is used. 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.

REFERENCES

[1] S Kaviarasan, K Hema Priya, K Gopinath :Semantic web usage mining techniques for predicting users' navigation requests Journal of Innovative Research in Computer and Communication Engineering.

[2] HongyuGao NorthwesternUniversity Evanston, IL, USA hygao@u.northwestern.edu, Jun Hu HuazhongUniv. OfSci. &Tech and Northwestern University, Detecting and Characterizing Social Spam Campaigns.

[3] S. Lee and J. Kim, Warning Bird: 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,Neurocomputing, http://dx.doi.org/10.1016/j.neucom.2015.02.047

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

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

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

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

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

[12] 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.

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

[14] 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.

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

[16] http://www.cs.bu.edu/fac/gkollios/ada05/LectNo te s/lect25- 05.pdf [17] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, Detecting spammers on Twitter, in Proc. CEAS, vol.6.2010.