Detection of Twitter Spam's using Machine Learning Algorithm

K. Jino Abisha[1], J.Roshan Nilofer[2], A.Silviya[3], Dr. S. Raja Ratna[4]
[1,2,3] *IV CSE, V V College of Engineering, Tisaiyanvilai, India*
[4]*Associate Professor, CSE, V V College of Engineering, Tisaiyanvilai, India*

Abstract

With the increased popularity of online social networks, spammers find these platforms easily accessible to trap users in malicious activities by posting spam messages. In this work, Twitter platform is taken and spam tweets detection is performed. To stop spammers, semi supervised learning is used to detect spam tweets in twitter. Thus, industries and researchers have applied different approaches to make spam free social network platform. Some of them are only based on user-based features while others are based on tweet based features only. To solve this issue, a framework has been proposed which takes the user and tweet based features along with the tweet text feature to classify the tweets. The benefit of using tweet text feature is that the spam tweets can be identified even if the spammer creates a new account which was not possible only with the user and tweet based features. The work has been evaluated with three different machine learning algorithms namely - Support Vector Machine, Neural Network, Random Forest. With Naive Bayes classifier, about 80% of accuracy is obtained.

Keywords - *Twitter, spam, supervised learning, support vector.*

I. INTRODUCTION

In the past few years, online social networks like Face book and Twitter have become increasingly prevailing platforms which are integral part of people's daily life. People spend lot of time in micro blogging websites to post their messages, share their Ideas and make friends around the world. Twitter is rated as the most popular social network among teenagers. However, exponential growth of Twitter also invites more unsolicited activities on this platform. Nowadays, 200 million users generate 400 million new tweets per day.

This rapid expansion of Twitter platform influences more number of spammers to generate spam tweets which contain malicious links that direct a user to external sites containing malware downloads, phishing, drug sales, or scams . These types of attacks not only interfere with the user experience but also damage the whole internet which may also possibly cause temporary shutdown of internet services all over the world.

II. LITERATURE SURVEY

This paper [1] addresses the task of detecting social bots in Twitter-like SNSs by developing a semi-supervised collective classification technique, called SocialBotHunter, which combines the social behavior of users and social interactions among them in a unified manner. SocialBotHunter relies on the homophily property, meaning love of the same. It takes a social graph of users and a small set of labeled legitimate users as input and then trains an OCSVM classifier by the labeled legitimate users to estimate the initial anomaly scores of unlabeled users. The results indicate that SocialBotHunter performs significantly better than previous social botnet detection techniques.

Erwin et al [2] discusses a different approach compared to previous research are the scope of Indonesian-language Twitter, crawling automatically for user and tweets data, as well as the addition of new features. Two features dimension are used, i.e., user-based and tweet-based. In this paper, detect Indonesian spammers on Twitter using four classification algorithms, namely Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression and J48. The results are confirmed for having better accuracy that of the existing. The highest accuracy of 93.67% is achieved using Logistic Regression.

Li et al [3] discusses the Laplacian score method, which is an unsupervised feature selection method, to obtain useful features. Based on the selected features, the semi-supervised ensemble learning is then used to train the detection model. Experimental results on the Twitter dataset shows the efficiency of the approach after feature selection. Moreover, the method remains high detection performance in the face of limited labeled data.

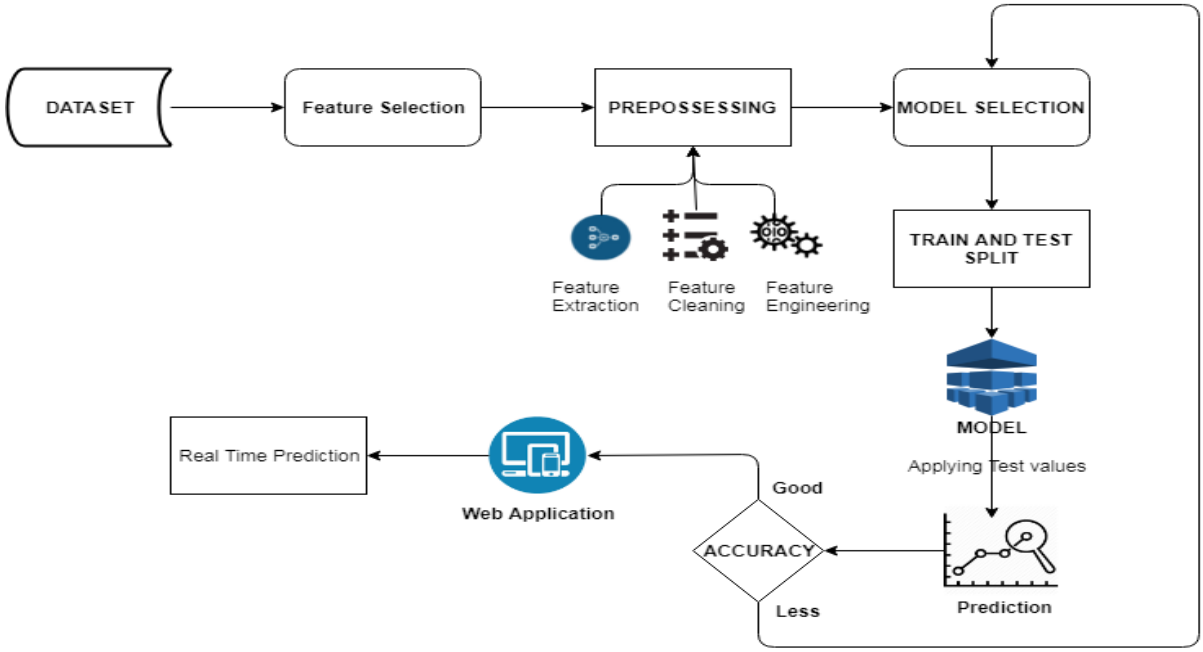

Fig.1 Spam detection process

Boyd et al [4] has done a lot of research to detect spam profiles in OSNs. This paper reviewes the existing techniques for detecting spam users in Twitter social network. Features for the detection of spammers could be user based or content based or both. Current study provides an overview of the methods, features used, detection rate and their limitations for detecting spam profiles mainly in Twitter.

Nguyen [5] proposed another form of deep learning, a linguistic attribute hierarchy, embedded with linguistic decision trees, for spam detection, and examine the effect of semantic attributes on the spam detection, represented by the linguistic attribute hierarchy. This approach not only efficiently tackle 'curse of dimensionality' in spam detection with massive attributes, but also improve the performance of spam detection when the semantic attributes are constructed to a proper hierarchy.

III. SPAM DETECTION PROCESS

The spam detection process consists of three modules as shown in Fig.1.

a) Analyzing data
b) Feature selection
c) Preprocessing
d) Model building
e) Accuracy

A. Analyzing data

The Dataset obtained from Kaggle with Costumer Complaint data machine learning group. For confidentiality the dataset is simply provided as 28 unlabeled columns. The data in the dataset are analyzed using algorithms.

B. Feature Selection

The features are extracted from the dataset. Some of the features are text based, vocabulary based, metadata based and so on. About 21 features are selected. The benefit of using those features based on their entropy score were able to reduce uncertainty in the prediction outcome. After collecting 20,000 labeled tweets, around 15,000 tweets are extracted.

C. Preprocessing

In preprocessing the tokenization of each message in the dataset takes place. Tokenization is the job of splitting up a message into pieces and removing the punctuation characters.

D. Model building

After the preprocessing, the model is build according to the features used in Semi-supervised learning, which uses labeled and unlabeled data. That is a great opportunity for those who can't afford labeling their data. The method allows us to significantly improve accuracy, because unlabeled data are used in the training set with a small amount of labeled data.

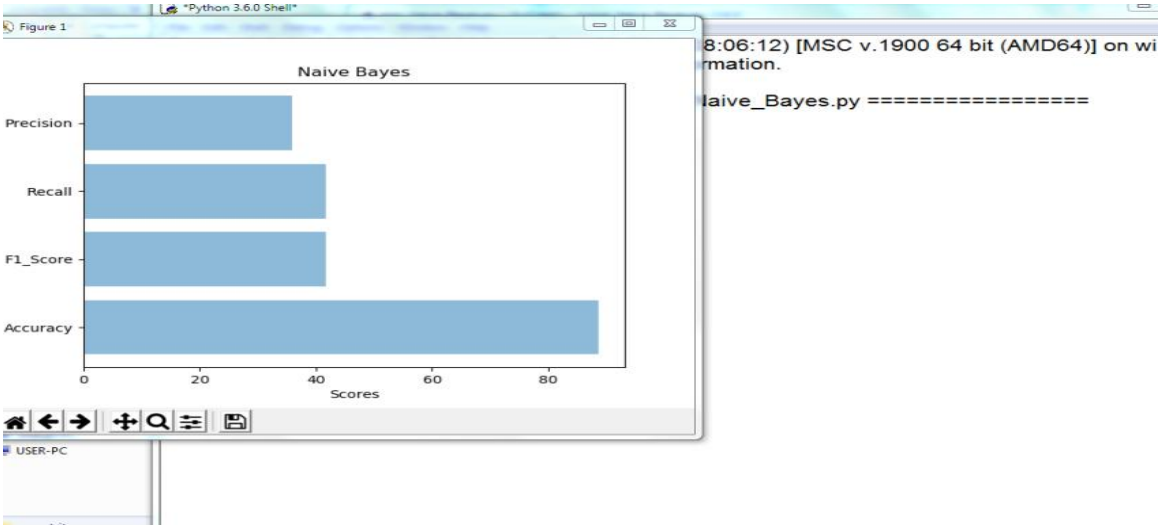

Fig. 2 Representation of Accuracy maintained by Naive Bayes

E. Naive Bayes

Naive Bayes is based on two assumptions. Firstly, all features in an entrance that needs to be classify are causative evenly in the decision (equally important). Secondly, all attributes are statistically self-determining, meaning that, knowing an attribute's value does not indicate whatever thing about other attributes' values which is not always true in practice. The process of classifying an instance is done by applying the Bayes rule for each class given the occurrence. In the fraud detection task, the following formula is calculated for each of the two classes (fraudulent and legitimate) and the class associated with the higher prospect is the predicted class for the instance.

IV. IMPLEMENTATION AND RESULTS

The result discusses two main parameters the ROC Curve and the confusion matrix. By looking at the Area Under the Curve, it is easy to determine whether the ROC curve is good or bad.

A. Accuracy in Navie Bayes

The Fig. 2 represents the accuracy maintained by Naive Bayes. Fraud detection is a binary classification assignment in which any contract will be predicted and labeled as a fraud or legit. The proposed classification techniques were tried for this task and their performances were compared. The following subsections briefly make clear the Naive Bayes, data set and metrics used for routine measure.

B. Confusion Matrix in Spam detection:

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. All the measures except AUC can be calculated by using left most four parameters. The below Table 1 discusses about the four parameters.

TABLE I
REPRESENTATION OF CONFUSION MATRIX

		Predicated Class	
		Class=yes	**Class=no**
Actual Class	Class =yes	True positive	False Negative
	Class =no	False positive	True negative

True positive and true negatives are the observations that are correctly predicted and therefore shown in green. False positives and false negatives has to be minimized so they are shown in red color. These terms are a bit confusing. So let's take each term one by one and understand it fully.

A. True Positives (TP)

These are the correctly predicted positive values which mean that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

B. True Negatives (TN)

These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing. False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

C. False Positives (FP)

When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

D. False Negatives (FN)

When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger

survived and predicted class tells you that passenger will die. Once the four parameters are identified then Accuracy, Precision, Recall and F1 score can be calculated. About 80% of accuracy is maintained in this proposed work.

After calculating metric score if the metric score is less, then it can't be used in real time and if the accuracy is good then it can be implemented in real world.

V. CONCLUSION

In the real world, spam tweet's feature keeps on changing in an unanticipated way. This problem is referred as Spam Drift. For classifying tweets as spam and non spam there are various techniques used. As Twitter API is available to all users, spammers may change their behavior over the time. This paper discusses semi supervised learning to detect spam tweets in twitter. It is tested with real-time tweet detection and it outperformed existing approach by 18%. It is observed in the dataset considered, 79% of spam tweets contain a malicious link. So, URL crawl mechanism is performed to detect Twitter spam. However, the concept takes up a classification algorithm and suggests various improvements that directly contribute to the advance of accuracy. The model has been tested in time period and may capture live streaming tweets by filtering through hash tags so perform immediate classification.

REFERENCES

[1] Dorri. A., Abadi, M., and Dadfarnia, M. (2018), "Social Bot Hunter: Botnet Detection in Twitter-Like Social Networking Services Using Semi-Supervised Collective Classification", IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, 2018.

[2] Erwin B. Setiawan, Dwi H. Widyantoro, and Kridanto Surendro, "Detecting Indonesian Spammer on Twitter", School of Electrical Engineering and Informatics, No 10, 2018.

[3] Li, W., Gao, M., Rong, W., Wen, J., Xiong, Q., and Ling, B, "Lssl-ssd: Social spammer detection with laplacian score and semi-supervised learning" International Conference on Knowledge Science, Engineering and Management (pp. 439-450), 2016.

[4] M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," Journal of Computer Mediat. Communication, vol. 13, no. 1, pp. 210– 230, Oct. 2007.

[5] H. Nguyen, "state of social media spam", Nexgate, Research Report, 2013.

[6] Z. Chu, I. Widjaja, and H. Wang, "Detecting social spam campaigns on Twitter," ACNS 2012.

[7] J. Zhang, R. Zhang, Y. Zhang, and G. Yan, "The rise of social botnets: Attacks and countermeasures," IEEE Transaction on Dependable Secure Computation.

[8] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spam bots: Evidence, theories and tools for the arms race," Proceedings in. 26th International Conference on World Wide Web Companion, Apr. 2017, pp. 963–972.

[9] S. Lee and J. Kim, "Warning bird: A near real-time detection system for suspicious URLs in twitter stream", IEEE Transaction Dependable Security Comput., vol. 10, pp. 183–195, 2013.

[10] Vishwarupe, M. Bedekar, M. Pande, and A. Hiwale, "Intelligent Twitter Spam Detection : A Hybrid Approach," Smart Trends System Security, ," Proceedings in. 26th International Conference on World Wide Web Companion, pp. 189–197, 2017.

[11] Verma, M., and Sofat, "Techniques to detect spammers in twitter-a survey", International Journal of Computer Applications, 2014.

[12] He, H., Watson, T., Maple, C., Mehnen, J., and Tiwari, A, "A new semantic attribute deep learning with a linguistic attribute hierarchy for spam detection" IEEE International Joint Conference on Neural Networks, pp. 3862-3869, 2017.

[13] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam : The Underground on 140 Characters or Less Categories and Subject Descriptors," Proceedings in 17th ACM Conf. Computation, Communication and Security, pp. 27– 37, 2010.

[14] J.Martinez-romo and L. Araujo, "Expert Systems with Applications Detecting malicious tweets in trending topics using a statistical analysis of language," Expert System Application, vol. 40, no. 8, pp. 2992–3000, 2013.

[15] A. H. Wang, "Don't follow me: Spam detection in Twitter," Proceedings in International Conference in Security and Cryptography, vol. 2010, pp. 1–10, 2010