IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

Spam Posting Account Detection in Twitter

Haritha Kadimuttath[1], Savyan P V[2]

M. Tech., Student, Computer Science and Engineering, GEC Palakkad, India[1]

Assistant Professor, Computer Science and Engineering, GEC Palakkad, India[2]

Abstract: Social media platforms are globally connected inter network through which users can interact with each others. Twitter is one of such social media platform. Because of the globally connected nature of social media platforms, once a user post something on social media it may reach out to everywhere around the world. This is the advantage and disadvantage of social media platforms. Some people use this feature to promote their products and needs, the legitimate users. Some of them use for spreading unwanted and illegitimate contents, the spam users. This project aims for detecting and classifying legitimate and illegitimate accounts on Twitter based on different user activities. The user activities include tweets of a user, friends count of user, followers count of user, list count of user etc. To build the model data of users is collected from Twitter API. And to label the dataset collected from twitter unsupervised machine learning algorithm like k-mean clustering technique is used for this project. And for the detection and classification of spam users this project use a machine learning model and this model performs better than other machine learning algorithms such as Random Forest, Decision Tree and Multinomial NB. And this work also capable to work on real time twitter data so that users can easily identify the spam users on real time.

Keywords: Machine Learning, Twitter API, Labelling, Classification

I. INTRODUCTION

Today's world is connected through internet and there are so many social media platforms working on internet. The social media platforms include facebook, whatsapp, instagram, twitter etc. Nowadays most people in the world use social media platforms for sharing anything to the world. Because the social media is connected, anything shared using this may reach out to any part of the world and it is freely available for its users. Due to the global connectivity of social media platforms many organisations and industries use these for their promotion purpose. This feature is misused by many others so as to promote unwanted or irrelevant contents. Such users are called fake users, spam users or simply spammers. So it is important identify such users from legitimate users. Twitter is one of such online platform in which this project is concentrates on. What is a spam message? A spam message is a message if (a) the receiver's personal identity and context are not considered because the message is equally applicable to many other potential receivers; and (b) the recipient has not verifiably granted intentional and explicit permission for that to be sent. This can also state this as in twitter any tweet that is highly irrelevant and useless to the user it is being sent to. This includes both the text within the tweet itself, and any web pages it may link to. From this definitions we can conclude that a user who sent such spam messages is a spam user. In the current scenario there may be millions of spam accounts that spread unwanted contents through twitter. The figure 1 shows accounts that are detected spam during the period september 2017 to may 2018. The graphical representation shows the tremendous growth of spam users on twitter. From this it is clear and evident that the spam accounts in twitter is growing very fastly and we need to get rid of it. And a study shows [1] one in every 200 social media messages and one in every 21 tweets is estimated to be spam. So the spam accounts.

Fig. 1 The rapid growth of spam accounts on twitter

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 6, June 2020

Most of the spam users use tweets to convey their intentions and while checking their accounts such users have some specific patterns. By identifying such user patterns it will be easier to identify spam users from the legitimate one. But manually identifying this is a difficult task so we can use a machine learning algorithm for the classification. The figure 2 shows the basic working of spam account classifier. So the main task is to identify the pattern of user activities.

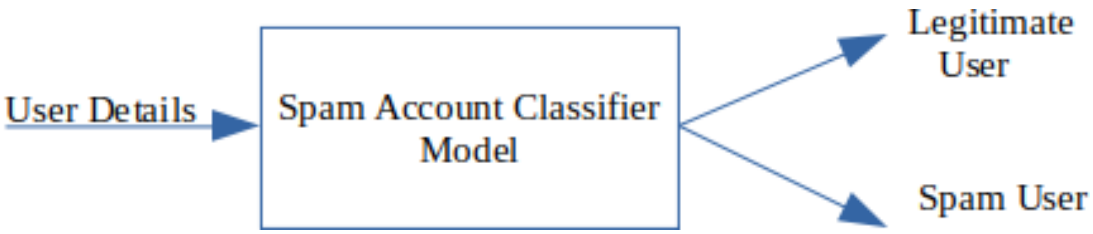

Fig. 2 Simple working of spam account classification model

A. Applications

1) Instant spam tweet identification
2) Real time spam account detection

The application of spam account classification in social media platforms include instant detection of spam accounts and real time identification of spam messages. The data from the twitter is also used for other areas of application such as sentiment analysis, product review etc., so the data should be credible for these studies. So it is important to avoid spam messages from these platforms for these areas of application. The section 2 includes various machine learning methods for spam account classification. Section 3 and 4 includes new machine learning model for the classification of spam accounts and Section 5 discussing about the comparison between existing models and new model.

II. DIFFERENT MODELS FOR SPAM ACCOUNT CLASSIFICATION

This section deals with various machine learning models currently used for spam account classification. These methods mainly uses existing machine learning model for the classification.

A. Methodology-1 Adewole et.al 2020

This framework [2] targets spam message and spam account detection using twitter and mobile data as test samples. Tweet can be embedded with entities such as hashtag, mention, and shortened urls. Spammers employ mention tool for target attack since the twitter featured a unidirectional user binding. Although twitter has introduced features to deactivate unsolicited mention, however, a majority of users on twitter still utilize default account settings. The visibility of a tweet on the network is increased through a process of retweeting. Retweeting a user's tweet has been identified as another strategy used by spammers to keep their accounts running . In addition, spammers' accounts exhibit automated posting patterns since there is a need for spammers to get across to a large number of users on the network. Even though twitter has become an important platform for real time communication, however, it has gone through several cases of abuses in the hands of social spammers. This is evidence in the rules introduced by twitter to suspended account with abusive behaviors.

B. Methodology-2 Concone et.al 2019

The model [3] proposed by Concone et al. classifies spam account using labeling technique. And for the labeling url is considered. They thought that spam users use their urls to identify the spam accounts. That is spam users use these malicious urls for their works. The first point to consider is the publication of malicious urls that direct to phishing sites or induce users to download unwanted software. Detecting such links is not simple because spammers adopt strategies that confuse the target url, thus deceiving the end user. For this reason, despite the possible counter measures, links are the easiest way to disseminate malicious contents. The labeling is done using url analysis, and similar tweets discovery.

C. Methodology-3 Alom et.al 2018

To detect spam accounts, this work [4] considered both graph based features and content based features. Graph based features include triangle count of user's network, the ratio of triangle count to the number of followers of a user, and the ratio of bi-directional links and content based features include unique url ratio, url to tweet ratio, average tweets per day of a user, and average likes per tweet of a user. To assess this detection method, they selected, from the popular social honeypot dataset, 325 Twitter accounts, where 168 are considered legitimate users and 157 spam users. They used several machine learning classification algorithms for distinguishing between spammers and non-spammers accounts. They implemented it using machine learning techniques such as SVM, KNN, RF, and Logistic Regression.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 6, June 2020

D. Methodology-4 Dutse et.al 2018

This approach [1] proposes a set of features that independent of past tweets, which are only available for a short time on twitter. They take into account features related to the users on twitter, their account information and their pairwise engagement with each other. That is content based, network based and twitter specific memes.

III. MACHINE LEARNING APPROACH FOR SPAM ACCOUNT CLASSIFICATION IN TWITTER

Machine learning models work better for classification related works than deep learning models. Mainly there are three kinds of machine learning approaches, supervised, unsupervised and reinforcement learning. This work use both supervised and unsupervised machine learning algorithms. Unsupervised learning is used when there is no labeled dataset. And supervised learning is used when there is a labeled dataset. The data collected from twitter is unlabeled so to label the dataset unsupervised learning technique is used. And for the classification machine learning technique is used. Figure 3 shows the basic working flow of proposed spam account classification model.

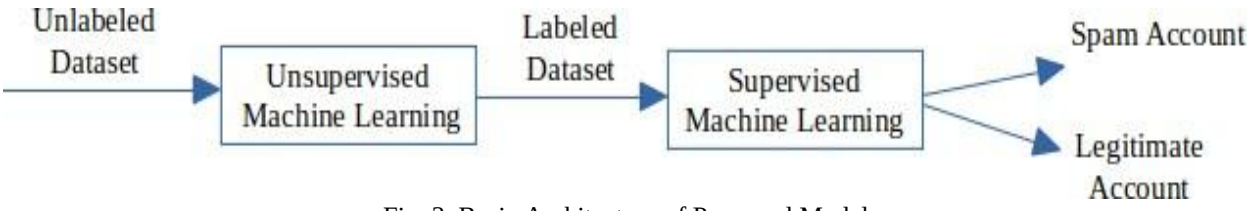

Fig. 3 Basic Architecture of Proposed Model

IV. SPAM ACCOUNT CLASSIFICATION IN TWITTER

Twitter is one of the widely used social media platform in which large amount of data produced everyday. From these this project classify spam accounts from the legitimate accounts. Figure 4 explains the detailed architecture of the model. The proposed model contains two parts label the dataset according to the spam tweets, and then detect spam accounts by analysing these labels and some other features. This is the better way than other models because they do not consider tweets for their spam account classification. For that this model first label the dataset based on the tweets, by using a unsupervised machine learning algorithm like k-mean. This labeled dataset is then used for the classification of accounts between spam and legitimate. For the classification the features selected based on network based and content based. The list of features are given in table 1. Using these features a model build so as to classify the accounts into spam and legitimate users.

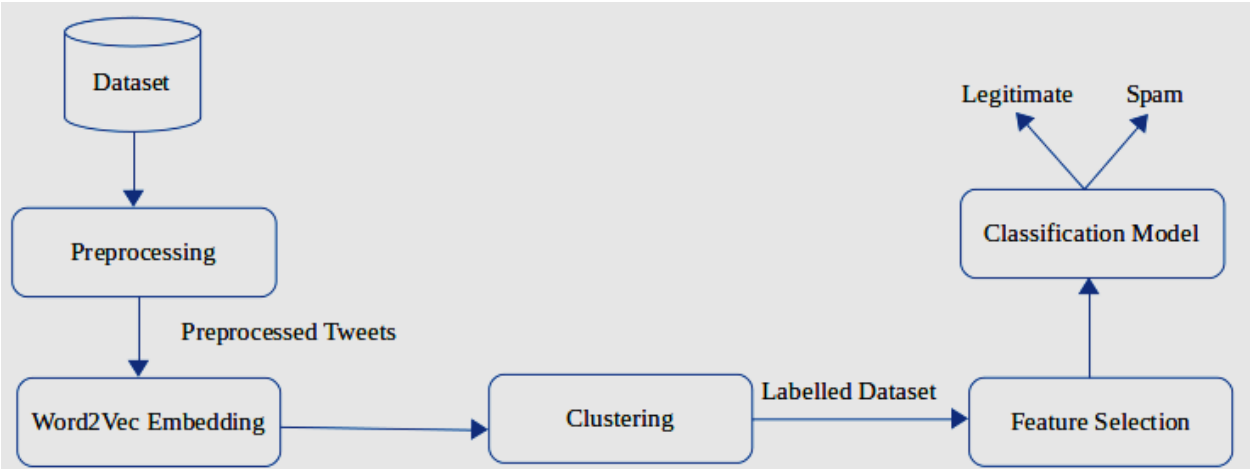

Fig. 4 System Overview

A. Labeling Module

The data required for this model is collected from the twitter API. The data collected from the twitter is not labeled, in order to build a classification model that need a labeled dataset. So the labeling is important. For the labeling process tweets are considered. So to make a labeled dataset, word2vec embedding [5] and k-mean algorithm [6] on tweets are used for the labeling process. The tweets are preprocessed [8] first and then it is embedded using word2vec mechanism, these word vectors are then fed to k-mean model to label the dataset. The preprocessing step includes non english word elimination, tokenisation, lemmatization, removal of stopwords and removal of punctuations. After the preprocessing, these preprocessed tweets were embedded and then used for the labeling purpose.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 6, June 2020

B. Classification Module

A new algorithm [7] is used for the classification purpose. The input to the classification module is the data obtained after labeling. These dataset contain features that useful for this model and some irrelevant features also. Feature selection is an important task to be performed. The features used for this project is mainly network based features, user profile based features and content based features. The features used for this project are specified in detail in the table 1. The model use bag of words (a set of spam words) mechanism on content based features such as screen name, user name, description and tweets to find these content based features contain any spam content on them. This are analysed using bag of word mechanism. Along with these content based features the network based features and user profile based features are also used for the model building. User profile based features contain information about the user and network based features give information about the users interconnection with other users. These features are fed in to the proposed model for the training process. 75% of the data is used for training and rest for testing. The newly build model takes input from user and classify them in to spam and legitimate accounts with very low false positive rate.

Table 1 Features Used

Network Based	**Content Based**	**Profile Based**
Friends Count	Description	Profile Image
Favourites Count	Tweets	Verified User
Followers count	Screen Name	Status Count
Friend Follower ratio	User Name	
List Count	Url	
Follower Retweet Count		

V. SPAM ACCOUNT CLASSIFICATION IN TWITTER

The overall comparison between the proposed model and the other machine learning models that implemented for this comparison are discussed here. The newly build model have better performance than the other existing machine learning model with the same features. And this model have very small false positive rate and high true positive rate compared to other models also. The table 2 shows the accuracy of new model with the existing machine learning models such as random forest, decision tree and multinomial naive bayes.

Table 2 Evaluation of Different Models

	Training Accuracy	**Testing Accuracy**
Proposed Model	0.9646	0.9385
Decision Tree	0.8824	0.8785
Random Forest	0.8252	0.7916
Multinomial NB	0.5421	0.5631

VI. CONCLUSION AND FUTURE SCOPE

This paper proposes a new model for spam account classification in twitter. The newly build model have better accuracy than other existing models such as random forest, decision tree and multinomial naive bayes. And this model also have high true positive rate and low false positive rate compared to these model. This model is also capable to work on real time spam account detection.

In future, the system can also modify in a way that, for labelling process use Latent Semantic Analysis (LSA). By using this it would be label the class through the semantic similarity. This will give better performance to the model.

ACKNOWLEDGMENT

First and foremost I wish to express my whole-hearted indebtedness to God Almighty for his gracious constant care and blessings showered over me for the successful completion of the project. Moreover, I express my deep gratitude to family, friends and teachers for their whole hearted support.

REFERENCES

[1]. I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on twitter," Neurocomputing, vol. 315, pp. 496–511, 2018.
[2]. K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spamaccount detection based on clustering and classification methods,"The Journalof Supercomputing, vol. 76, no. 7, pp. 4802–4837, 2020.
[3]. F. Concone, G. L. Re, M. Morana, and C. Ruocco, "Twitter spam account detection by effective labeling.," in ITASEC, 2019.

[4]. Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on twitter," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1191–1198, IEEE, 2018.
[5]. B.Salehi, P.Cook, & T.Baldwin, "A word embedding approach to predicting the compositionality of multiword expression," in Proceeding of the 2015 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 977–983, 2015.
[6]. A. K. Jain, "Data clustering: 50 years beyond k-means,"Pattern recognitionletters, vol. 31, no. 8, pp. 651–666, 2010.
[7]. J. Brownlee,Master Machine Learning Algorithms: discover how they work andimplement them from scratch. Machine Learning Mastery, 2016.
[8]. C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp. 55–60, 2014.

BIOGRAPHY

Haritha Kadimuttath earned her B. Tech., degree from CUSAT, Kerala, India. And currently pursuing M. Tech. in Computational Linguistics from GEC Palakkad, APJ Abdul Kalam Technological University, Kerala, India respectively. Her area of interest includes machine learning and deep learning.