A Multi-classifier Framework for Detecting Spam and Fake Spam Messages in Twitter

R.Jeberson Retna Raj, Senduru Srinivasulu, Aldrin Ashutosh
Department of Information Technology, School of Computing,
Sathyabama Institute of Science and Technology, Chennai, India.
Jebersonretnarajr@gmail.com

***Abstract*—Social media plays vital role among the user communities for social gathering, entertainment, communication, sharing knowledge so on. Twitter is one such network to connect millions of users to share information. Nowadays, there are humpteen numbers of users using social media for social engagements. Due to the fact that wide publicity of individuals and products get viral in social media, everyone wish to use social media as a platform to promote their product. Furthermore, large number of people relies on social media contents to take decisions. Twitter is one of the social media platforms to post the media contents by the user. Spammers are illegal users intrude the twitter account and send the duplicate messages to promote advertisement, phishing, scam and personal blogs etc. In this paper, a novel spam detection mechanism is introduced to detect the suspicious users on twitter. The system has been designed such a way that it initially set with semi-supervised at the tweet level and update into supervised level for learning the input tweets to detect the spammers. The proposed system will also identify the type of spammers and will also remove duplicate tweets. We have applied with multi-classifier algorithms like naïve Bayesian, K-Nearest neighbor and Random forest into twitter data set and the performance is compared. The experimental result shows very promising results.**

Keywords—Twitter Spam, Multiclassifier, Classification;

I. INTRODUCTION

Social media is one of the platforms used by large number of users for learning, entertainment, promoting advertisement and social engagement. Through social media one can share the messages or information to millions of users at a time. Survey shows that individuals spent as an average of 144 minutes per day on social media since 2014 to 2019. In 2019, a total of 4.4 billion internet users worldwide and every minutes there are 4,79000 tweets generated. With the advent of internet and advancement of technology IoT technology integrated technology with the smart devices which generates large volume of data. In 2020, it is expected that 44 zetta byte of data will be generated by various sources connected with internet. Nowadays social media is an important source for individuals and corporate for taking business decisions, report, and opinion and so on. Data is an important asset and with the help of the available technology, one can extract the essence of the data. Due to fact that a sizable amount of gain and wide publicity can be done with the social media, social spammers making use of the network and spread fake and false information to the media users. Twitter is the one of the most wanted media network used by its users. In this paper, we introduced a framework which used to process thousands of tweets per minute and able to detect the spammers. Furthermore the system deletes the spam messages. The proposed system equipped with semi-supervise framework for spam detection and multi-classifier algorithm for differentiating the spam messages. The multi classifier algorithms which includes Naive Bayesian classifier, K-Nearest, Random Forest and Decision tree is applied with the data set and the accuracy is compared. The multiple classifier algorithms are efficiently detecting the spam messages. Furthermore, the system identifies duplicate or redundant spam messages. The similarities of the tweets are identified and categorized it namely leg data, spam data and total data. The legdata describes the particular person and the spam data describes how many of those posts are spams of that particular person. The total data is used to count the overall twitter data tweeted by everyone. The proposed spam detection system used to learn the tweets and the activities, and accurately classifying the new data inputs.

II. LITERATURE REVIEW

Twitter was launched on 2006, soon after getting launched it has became a hotspot for scamming, phishing by means of spamming in tweets. In today's time spam in twitter is a common practice performed by spammers. The spam can be in various form for promoting advertisements to works which are illegitimate. The spammers may tweet tweets which they had previously done over and over again. Efforts are made to stop people from spamming. It becomes difficult to handle such kind of situations as the spammers find an alternative to get pass the security which causes an disastrous intrusion. The intrusion can also be in the form of spammer tweeting duplicate or near duplicate tweets in other words tweets which has similarity to some other tweets. Therefore, an anti spam system must be developed to tackle such kind of situations. So, it must be necessary to find the spams which are tweeted live. This can be done using by creating an account as developer API which will in turn return or retrieve the live data. Through this one can understand the pattern of how people are trying to spam in real time. There are several methods to find how frequent the spammer is by getting his data by knowing how time the

978-1-7281-4976-9/20/$31.00 ©2020 IEEE
DOI: 10.1109/CSNT.2020.49

person uses the same URL, this is done by getting AvgURLcount. How many times the person is mentioning the same tweet by using AvgMentioncount. Also check if the spammers is re-tweeting or not. This can be done by the identifying the average retweets, this is done by using Avgretweets. Also it is used to get average favorite tweets by using Avgfavcount.

Wang et al proposed K-L divergence method used for extract the concept in the distributed pattern of spam messages and Multi Scale Drift Detection Test (MDDT) used to identify drift detection in the spam messages [1]. Sedhai et al presented a spam detection model for detecting spam messages in twitter. The real time processing of tweets are processed by spam detection module and the batch processing mode works in offline processing of twitter data. The framework equipped with four lightweight detectors are used to define the malicious contents in the twitter stream. These detector used to detects spam url, duplicate tweets etc [2]. Chen et al presented a machine learning algorithm to classify the drifted spam messages. Random Forest algorithm is applied with the tweets for classifying the spam messages and non-spam message contents [3]. Madisetty et al proposed an ensemble based multi layer deep neural network CNN algorithm is used to detect the spam in the twitter stream. Each CNN uses five different word embedding techniques used to train the twitter data. The method combines with deep learning algorithm and feature based models to detect the spam data [4]. In [5], a spam detection mechanism is introduced to classifying spam messages. Naïve bayers classifier algorithm is used to classifying the tweets. The input tweets are divided into three micro clusters and checking the similarity threshold exceeds the maximum then group them into a micro clusters and less than the cluster is grouped into other cluster. In [6], an hybrid approach is introduced to detect the spam messages and categorized into three different categories with the specific features such as metadata, content and interaction based features. In [7], a twitter spam detection mechanism is introduced to identify spam in trending topics, fake users, fake content and spam based on url. In [8], a microblog content extraction method introduced to extract 5WTAGs based on the hash tags such as What, When, Who, where and hoW. In [9], a real time URL based spam filtering mechanism in Twitter has been introduced. In [10] and [11], the IoT and Bigdata analytics helps in social media

III. PROPOSED METHOD

The proposed spam and fake spam message detection framework contains four main detector modules such as hashtag-based features, content-based features, user-based features, and domain-based features. The architecture of the proposed system is shown in figure 1.

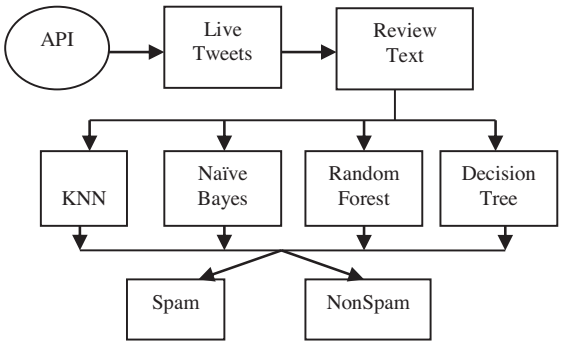

Figure 1. System Architecture

The system trained with black listed spamming domain and tested accordingly. The four detectors successfully classify the spam and non-spam tweets in the tweet window. Using the semi supervised method the required information is updated periodically based on the previous tweet window experience.

A. Duplicate tweet

This module will check if the tweets were tweeted earlier too. In other word, it will cross check if the tweets are duplicated or not. Here hashing can be used to remove duplicate tweets. This is done by using drop duplicate method. It will consider tweets as near duplicate tweets if one tweet show signs of similarity with any other tweets.

B. Multi classifiers

Here, multiple classifiers are used. This is used to increase the accuracy of system. In this each different classifier are used to satisfy the above situations based on the accuracy which provides accurate and high precision results. Several classifiers are used, namely, Naive Bayesian classifier, K-Nearest Neighbor, Random Forest, Decision tree. Among these, Random Forest will provide most accurate result and can be used further. The four classifiers use different classification techniques, i.e., generative, discriminative, and decision tree-based classification models. A full spectrum of features is extracted to represent each tweet. Features for tweet representation, the features include hashtag-based features, content-based features, user-based features, and domain-based features. Other than this, it will define several functions which were discussed earlier those are AvgHashtag which is used to find each person spamming. It will take the mean value of every month. AvgURL is used to find the mean/average URL used in a month. In other words, how many time each one URL is used. AvgMention is used to find the average times the user is mentioning or commenting a specific spam. AvgReTweet – this is used to describe average time the user is retweeting the same content. AvgFavCount is used to find his average favorite tweet over span of 30 days. Some of the features used are narrated here.:

C. *User-Behavioral (UB) based features:*

This feature is based on the understanding of how the spammer communicates in terms of wordings. As it is based on each individual user language, it contains two specialties. The first is burstiness of reviews written by a single user. The burstiness is nothing but a characteristics of communication which involves data to be transmitted in bursts, basically bursts is used to send data, large data particularly, in a short period of time which is usually triggered as a result of threshold being reached. Another noteworthy part of user behavioral feature is negative ratio has given to different business by average of users.

D. *Review-Linguistic(RL) based features:*

First and foremost, it is totally not based on metadata which means it doesn't contain any set of data that describes and give away information about it. It is actually based on review text where the data is extracted from the text itself. It also showcase some prominent features like Ratio of exclamation, first personal pronoun, and sentences containing 'res'.

E. *User-Linguistic(UL) based features:*

It is based on each individual user. It generalizes all the review written by the specific user. It is used to retrieve the live tweets from twitter through twitter developer API.

IV. EXPERIMENTS AND DISCUSSIONS

We used more than 10,000 tweets obtained through the dataset . The tweets are labeled as spam and ham in all tweets and the remaining portion of tweet is labeled as unknown. The tweets which are labeled as unknown due to the fact that they are not able determine their labels using manual observation. The twitter data is is classified using the four machine learning algorithms such as KNN, Random Forest, Bayesian and Decision Tree algorithms. The performance of the algorithm is compared with the metrics precision, recall, F-measure and accuracy. The proposed system considers the spam class as positive and non-spam class as negative. The system incorporates with Top-30 words of user-based features in the tweet text so that the system predicts the spam contents. This is helps to fine tune the system to identify spam contents and mitigate the loss due to spam. This property contributes to detect spam messages in real time by assigning 50% for training set and testing set.

A. *KNN algorithm*

The KNN algorithm used to find the distance between the new sample of training cases and find the k-closest one spam tweet to another. The algorithm further search the next closest spam tweet from the dataset. The system partitioning into training and validation data with the initial seed value samples. The following pseudo code can be used:

set.seed(101) index = createDataPartition(dataset PARAMETER)
train = data1[index]
validation = data1[-index]

the confusion matrix is used to check the dimensions of training and validation sets. The algorithm returns with the accuracy of 84% spam data in the tweet corpus. Figure 2 shows the performance of KNN algorithm

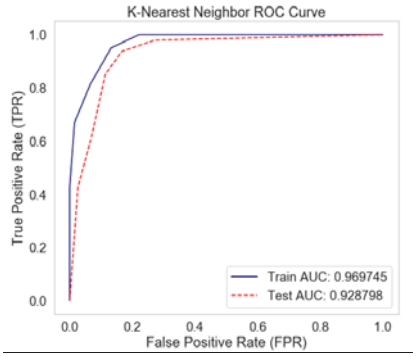

Figure 2: Performance of KNN algorithm

Here, it is computed by making use of true positive value against the fast positive rate. This provides better accuracy of 96% in training and 92% in testing.

B. *Decision Tree algorithm:*

The steps involved for Decision Tree algorithm is as follows:

- It begins with a dataset assume S
- It will calculate the entropy and information gain by iterating the algorithm
- Followed by it collects the sets which has the largest information gain or smallest entropy.
- It will then split S into spams and non-spam
- The algorithm continue to recur on each subset
- The training and testing accuracy obtained is 0.89 and 0.90 respectively

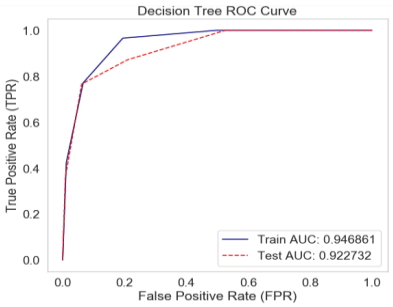

Figure 3: performance of Decision Tree algorithm

Figure 3 shows the performance of Decision Tree algorithm. Here, it gives 94 and 92 percentile accuracy for training and testing respectively

C. *Naive Bayesian algorithm:*

Let the spam tweet as spamt and word term as Wordt. The steps involved for Naïve Bayesian algorithm is as follows:

- Probability required for calculating afferent word in twitter. Formula may something look this:
- P(Spamt/Wordt)=P{(wordt/Spamt)*P(spamt)}/P(wordt)

- Create train test split
- Instating multinomial naive bayes classifier
- The accuracy of the dataset is 0.69 which is considered the lowest

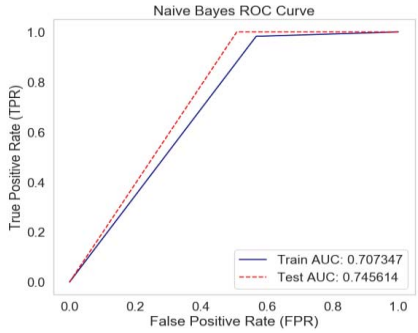

Figure 4: Performance of Naïve Bayes algorithm

Figure 4 presents the True Positive Rate against False Positive Rate, which in turn provides us 70% accuracy in training and 74% in testing

D. Random Forest algorithm:

The followings are the steps involve in Random Forest algorithm:

- It will extract the data from the review text.
- It instantiate random forest algorithm by training the dataset
- It will create an object rf and fit method is used for training and testing
  - rf = rf.fit(X_train, y_train)
  - y_pred_train = rf.predict(X_train)
  - y_pred_test = rf.predict(X_test)
- It will make predictions as well as predict probabilities to calculate the roc AUC for the spam detection.
- The training and testing accuracy obtained is 0.98 and 0.93 respectively.

Here, it is computed using True Positive Rate against False Positive Rate, which in turn provides us 70% accuracy in training and 74% in testing. Figure 5 shows the performance.

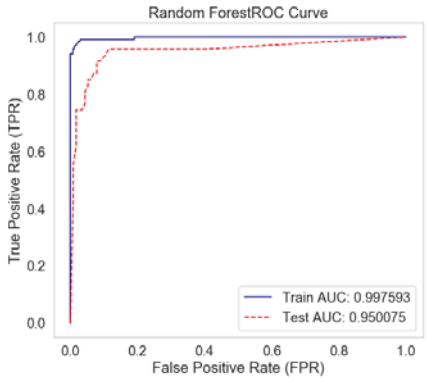

Figure 5 : perfromance of Random Forest algorithm

Figure 6 shows the overall structure, which combines the entire figure and illustrated in one. The test samples are experimented and the accuracy is shown.

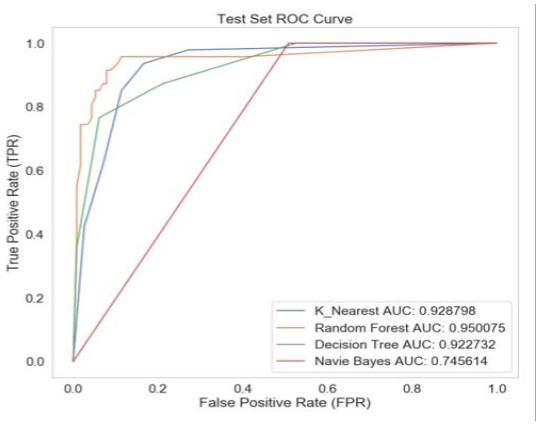

Figure 6: performance Comparison of algorithms

It clearly illustrates that random forest provides the best accuracy in terms of other classifiers both in training and testing itself. Confusion matrix is used to detect if there is any spam in the text, this is done by estimating the true predicted value if it is equal to actual value moreover if the predicted value is not equal to the actual value it will show as false values. It basically computes the value in a table with true and false in x-direction similarly true and false in y-direction. The classification report is obtained through confusion matrix as it contains true positive value(TP), true negative value(TN), false positive value(FP), false negative value(FN).

Precision: It is used to check how much prediction it turned out to be true from the list of positive tweets. It is not necessary the values must be as high as recall. The Tru Positive (TP), False Positive(FP) and False Negative is computed.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

Recall: In this, how much prediction is done perfectly from the given tweets. It is necessary the value must be high.

$$\text{Re}call = \frac{TP}{TP + FN} \tag{2}$$

F-score: It is also called as false measure. It is a evaluation measure to detect spams. F-score helps to measure recall and precision at the same time. This is done because it become very tough for comparing models with high precision and low recall or vice versa, that why F-score/False measure is used.

$$F - Measure = \frac{2 * recall * precision}{recall + precision} \tag{3}$$

V. CONCLUSION

In this paper, a novel framework for detecting spam tweets has been presented. The proposed system extracts the live tweets successfully through TwitterAPI. The system successfully extracts the data from the lists of review text. Multi-classifiers algorithms such as KNN, Naïve Bayes, Random Forest and Decision tree algorithms are applied with the dataset and the performance is compared. The multiclassifier algorithms successfully classify the tweets into spam and not spam tweets respectively. Among the algorithm with this dataset, Random classifier algorithm showcased with highest accuracy. The prediction mechanism successfully identifies with the binary classification of spam messages and no spam messages. Furthermore, the identified spam messages can be deleted.

VI. REFERENCES

[1] X. Wang, Q. Kang, J. An and M. Zhou, "Drifted Twitter Spam Classification Using Multiscale Detection Test on K-L Divergence," in *IEEE Access*, vol. 7, pp. 108384-108394, 2019.

[2] S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," in IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169-175, March 2018.

[3] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou and G. Min, "Statistical Features-Based Real-Time Detection of Drifted Twitter Spam," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, pp. 914-925, April 2017.

[4] S. Madisetty and M. S. Desarkar, "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," in IEEE Transactions on Computational Social Systems, vol. 5, no. 4, pp. 973-984, Dec. 2018.

[5] H. Tajalizadeh and R. Boostani, "A Novel Stream Clustering Framework for Spam Detection in Twitter," in IEEE Transactions on Computational Social Systems, vol. 6, no. 3, pp. 525-534, June 2019.

[6] M. Fazil and M. Abulaish, "A Hybrid Approach for Detecting Automated Spammers in Twitter," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 11, pp. 2707-2719, Nov. 2018.

[7] F. Masood et al., "Spammer Detection and Fake User Identification on Social Networks," in IEEE Access, vol. 7, pp. 68140-68152, 2019.

[8] Z. Zhao et al., "Modeling Chinese microblogs with five Ws for topic hashtags extraction," in Tsinghua Science and Technology, vol. 22, no. 2, pp. 135-148, April 2017.

[9] 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. Security Privacy, 2011.

[10] Armentano R, Bhadoria RS, Chatterjee P, Deka GC, editors. The Internet of Things: Foundation for Smart Cities, EHealth, and Ubiquitous Computing. CRC Press; 2017.

[11] Mazumder RS, Bhadoria RS, Deka GC. Distributed Computing in Big Data Analytics. InCon-cepts, Technologies and Applications 2017. Springer.