International Journal of Engineering and Techniques - Volume 6 Issue2 March- April 2020

Twitter Spam Classification using Machine Learning Techniques

Geetanjali Sharma[1], S.Samyuktha[2],Ishita Dhar[3], Golda Dilip[4]
Department of Computer Science and Engineering(4th Year)[1,2,3]
Associate Professor, SRM Institute of Science and Technology, Chennai, India[4]
SRM Institute of Science and Technology, Vadapalani, Chennai
Email:{geetanjali1298@gmail.com}
{samyuktha.suroju@gmail.com}
{ishitadhar27@gmail.com}
{goldadilip@gmail.com}

Abstract **-Stream clustering methods and mechanisms are used to categorize between spam and non-spam tweets. These methods make the assumption of considering the neighbouring clusters are symmetric and classify on that basis. However, this assumption is not entirely correct because the clusters have asymmetric distribution and might not be micro in size. Here, incremental naive bayes classifiers can be used for the microclusters whose population exceeds a certain level . This paper will further focus on applying the algorithms like Decision tree classifier, Support Vector classifier, Random Forest Classifier, Naive Bayes classifier and K-neighbouring algorithm to the dataset .Then, the performance of the algorithms will be tested on the basis of precision, recall, accuracy and F1 measure. The compared results further aid in proving that naive Bayes has the chance of being a better algorithm to classify spam and non-spam tweets**.

Index Terms **- Naive Bayes, machinelearning, spam and non-spam**

I. INTRODUCTION

Online Social Networks(OSNs) like twitter, instagram and facebook have seen tremendous growth in the industry. Thus, they have become very popular in a short span of time. Due to the popularity, these sites have become vulnerable to spammers. The latest statistics have shown that there are upto 200 million twitter members and approximately 400 million tweets are generated per day. The spam tweets include advertising messages, spreading of malware links through URLs, phishing attacks and frauds done financially. The common indicator of spam tweets are “hashtags”, shortened URLs and “mentions” though it is not always certain that the tweets containing these indicators are spam. Thus spam

filtering is a challenge to perform. As per the character limitation policy of twitter, there is a limit to the number of characters in a tweet because of which URLs are shortened. Spammers take advantage of these shortened URLs to spread spam tweets and malware links. It is not only the message content that needs to be monitored but also the behaviour of the user. For instance, Tweets sent to a large number of users which are greater than the connected people of that particular user, should be suspected as spam. However. Spammers send the spam tweets in batches to escape detection. They generally use fake trending hashtags to lure the users.

In different OSNs, spams are transferred mostly through the stream of messages shared by the user. Machine learning algorithms which are used in spam detection aids in identifying these spam messages. These detectors are constantly updated into better versions by incorporating supervised and as well as unsupervised methods. These methods are tailored and specialised to detect spams effectively. Despite supervised methods giving better results than the unsupervised methods , the cost of labelling large dataset prevents the spam filtering to be applicable. Classifier based approach is given to unravel the detection of spam messages. A classification model is an especially supported machine learning algorithm which provides the output within the sort of binary value. The feature extraction is a vital phase of the project to feature more benefits to the system. To tackle the issue of massive data, Naive Bayes is helpful. This algorithm has the capability of handling large dataset as well as streaming dataset. This algorithm works on Bayes Theorem of probability which detects probability of unknown dataset.

The remainder of the paper is as follows:

Section II describes the related work. Section III shows the design of the system. Section IV describes the methodology and working of the model.Theresult is discussed and presented in Section V. Conclusion is in Section VI with a brief description of future scope.

II. RELATED WORK

Various algorithms are used to differentiate the spam tweets from the genuine ones. Algorithms are compared and in some cases the performance is evaluated.

Shradha Hirve and SwarupaKamble [1] proposed a twitter spammer detection which would take a particular hashtag from any twitter's tweets. Since each hashtag contains thousands of comments and are constantly increasing. They handled this issue by using twitter4j API and performed preprocessing by discarding the hash symbols , quotes and through spam analysis

URL,Unsolicited Mentions (UIMn), Number Of Unique Mentions (NuMn). Google safe browsing API and Duplicate Domain Names (DuDn) techniques . However, increasing the training data does not aid in detecting twitter spam. They evaluated that the accuracy of detecting spam by the classifier decreased when used in a real-world scenario. They also concluded that spam tweets can be detected precisely if they are continuously sampled instead using selected tweets.

Shinde Asha Ashokrao and Shital Y. Gaikwad [2] proposed a method for filtering spam for social networks using automatic learning and classification techniques in which tweets are found in streaming and the Twitter provides access to developers and researchers for Streaming API for in order to access public tweets in real time. Since there was a gap for no evaluation of the continuous learning performance was taken, which was based on machine learning. They filled the gap by evaluating the performance based on three aspects, characteristics , data and model. The detection was converted to a binary classified problem and can be solved by ML algorithms. They listed the effect of different factors for performance of spam detection which were size for training Data, Data Sampling ,for data which was related to Time and as well as for semi-supervised learning algorithms.SurendraSedhai and Aixin Sun et al [3]proposed a framework that contains main modules, one module detects spam tweets while the other module is updated in batch mode. It consists of four detection which are light-weighted (i) Black listed domain detection to detect the tweet which contain blacklisted URLs (ii) near-duplicate
detector to identify the similar of other pre-labeled tweets (iii) ham detector that label tweets that are posted by trusted genuine users and do not consist of spam (iv)The remaining tweets and detected by multi-classifier.However, other features also should be observed when identifying the spammer.Miss. SalkeBhagyashri A. and Miss. Phad Kanchan R. et al [4] proposed that to accurately differentiate between spam and non spam featured extraction is crucial. Therefore, feature extraction becomes the important step. The system had 600 public tweets which were evaluated to identify the spammer and categories spam and non spam messages.

Faiza Masood and Ghana Ammad et al [5] focussed on the fake user identification on social networks. They observed the fake user and encourage services and websites that effect the legitimate users and harm resource consumption. This leads to the spread of harmful content. They focussed on detecting fake content , spam URLs, fake users and trending topics which are vulnerable to spammers. The presented techniques work on features such as content features, user features, time and graph features.

III. SYSTEM DESIGN

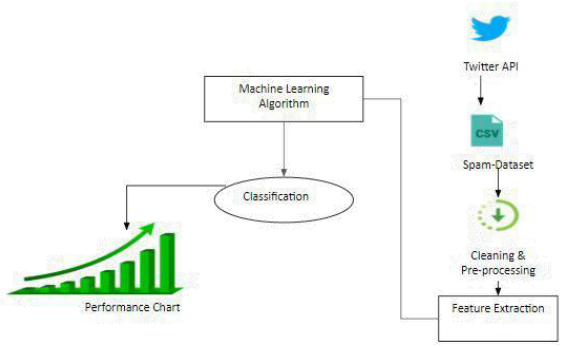


Fig. 1. Architectural diagram

Fig.1 demonstrates the process involved in developing the model. The diagram represents the key steps involved in the development of the proposed model. The sequence of operations like data processing , data cleaning and feature extraction takes place and in the end the classification is performed.

IV. METHODOLOGY

A. Data Collection

The dataset being used has both spam and non-spam tweets which were obtained from Kaggle. To differentiate between the tweets having both spam and non-spam. We need to identify the words which have the most probability of being present in a spam tweet. The data set is in a .csv file format which consists of 5573 tweets. These tweets need to be accessed and tokenized to gain the words from tweets and from non-spam tweets.Then this dataset is divided for training and testing randomly. 15% of the dataset is kept for testing while the remaining is used for training. Then, the dataset is ready to be analysed.

B. Machine Learning mechanism

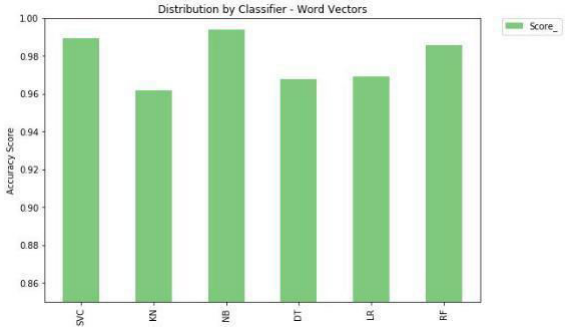
The dataset will be evaluated by using 6 machine learning classifiers which are Decision Tree Classifier, Support Vector Classifier, Random Forest Classifier,Naive Bayes Classifier, Linear Regression, K-neighbour Algorithm. Turn by turn the algorithms will classify the dataset into spam and non-spam.

C. Performance Statistics

The algorithm will be compared on the

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$



basis of their accuracy and f1 measure. This will be viewed in the form of a histogram.

Fig. 2. Histogram to compare performance of all the algorithms on the basis of accuracy.

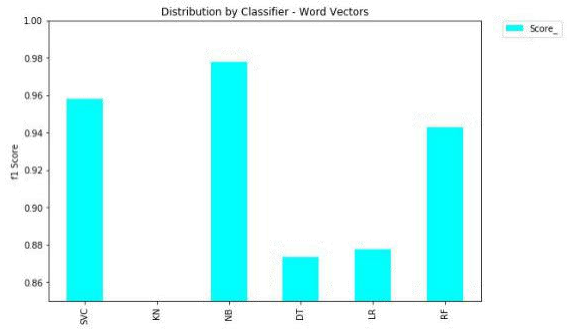


Fig. 3. The histogram which compares the algorithm on the basis of F1 measure.

D.User interface

The user will be able to register and login and then the user enters the tweet he/she is suspicious of being spam and the result will be calculated whether the tweet is spam or non-spam.

Fig. 4. User Interface that shows that the tweet is not spam.

Fig. 5. User Interface that shows that the tweet is spam.

V. Results

In this project, the content-based features and the user based features were compared. The performance of the algorithms on the dataset were also evaluated and their accuracy was compared. The algorithms that are compared, visualizes their accuracy and their performance on classification of non-spam and spam tweets. The algorithms we compare with are Decision Trees Classifier, Support Vector Classifier, Random Forest Classifier, Naive Bayes Classifier, Linear Regression, K-neighbour Algorithm. Among the six classifiers we evaluated,Our result shows that the Naïve Bayes classifier produces the best results. From the Fig .3 Our spam detector can achieve 0.97757 F-measure using the Naïve Bayes classifier.

VI. CONCLUSION AND FUTURE WORK

The Twitter popularity furthermore attracts the spammers and spammers send insignificant tweets to the users to lure them into visiting the websites which are malicious for traditional users. So for stopping spammers, researchers have planned different kind of mechanisms. From recent works the main focus is on applying machine learning techniques into the spam detection and the algorithms are been compared and visualized their

accuracy and their performance on classification of non spam and spam.

From our evaluation, we have seen that the classifier's ability of detecting Twitter spam reduced in a real-world scenario because unbiased data brings the bias data and also recognizes that the discretization of features was the significant preprocess to Machine Learning based spam detection. And secondly, by increasing the training data cannot bring additional benefits for detecting spam in a twitter after an appropriate number of training samples. We should try to bring more particular features or better model to further improve spam detection rate. Thirdly, classifiers can determine additional spam tweets when they are sampled repeatedly more than the randomly selected tweets. From the third perspective, we have figured out the reasons why classifier's performance decreased when training and as well as testing the data were on different days from three perspectives . We came to the final part that the performance decreases due to the fact that the features changes for the distribution of next days datasets, but the distribution of trainingdatasets remain constant. This problem will be there in streaming spam tweets detection, as the latest tweets are coming in the appearance of streams, but the training dataset wasupdated. In future, we will work on this problem.

REFERENCES

[1] Shradha Hirve, SwarupaKamble, “Twitter Spam Detection”, International Journal of Engineering Science and computing,vol 6 issue 10,Oct 2016.

[2]Shinde Asha Ashokrao, Shital Y. Gaikwad, “Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection”,International Journal of Innovative Research in Computer and Communication engineering, vol.5 Issue 1, Jan 2017, doi: 10.15680/IJICCE.2017.0501036

[3] Surendra Sedhai , Aixin Sun, “Semi-Supervised Spam Detection in Twitter Stream”,IEEE Transactions on Computational Social Systems (Volume: 5 , Issue: 1 , March 2018),doi:10.1109/TCSS.2017.2773581

[4] Surendra Sedhai ,Aixin Sun, “Spam Tweet Detection using Machine Learning Spproach”,International Journal of Advance Research and Innovative Ideas in Education (Volume: 4 , Issue: 3 , 2018),IJARIIE-ISSN(O)-2395-4396

[5] Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas ,Hasan Ali Khattak ,Ikram UdDin,MohsenGuizani ,Mansour Zuair, “Spammer Detection and Fake User Identification on Social Networks”,IEEE Access (Volume: 7),doi: 10.1109/ACCESS.2019.2918196