Journal of Advanced Computer Applications
vol-01, issue-01, December, 2024

ISSN No.: XXXX-XXXX
Pages 50-59

JOURNAL OF ADVANCED COMPUTER APPLICATIONS

ISSN: XXXX-XXXX (Online)
Vol: 01 **Issue:** 01
Contents available at: https://www.swamivivekanandauniversity.ac.in/jaca/

Machine Learning Approaches for detecting Spams in Tweets

Chayan Paul[1]*

[1] *Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore*
**chayanp@svu.ac.in*

ABSTRACT

During the last few years, the popularity of social networking sites has increased manifold. Microblogging sites like twitter became popular platform because of various at-tractive characteristics like direct connection with celebrities, sport persons, technocrats, businessmen, sharing and getting breaking news almost real time and so on. Because of the unprecedented growth in the user base of the network, there has been an immense growth in the number of spam messages in twitter also. Spams are un-wanted messages which are sent to multiple users in bulk predominantly for commercial promotional activities. Spams can be infuriating at times, as this may overload one's timeline and may hinder in getting the real messages. Also, spams may carry malicious links which may lead to embarrassing situations for the users. Because of the consequences of these issues, detection of spam becomes an important issue to deal with. During last few years, this problem has attracted considerable attentions from re-searchers and there have been some useful approaches. This paper takes an open-source data set and builds a collection of machine learning models to find the most effective algorithm that can be used to detect spams in a database of tweets. Algorithms namely XGBoost, AdaBoost, Random Forest and Decision Tree, were implemented on the selected data set for detecting the spam messages. This study was able to achieve an accuracy of 99.2% i.e., maximum out of all the classifiers we evaluated.

Keyword: Machine Learning Algorithms, Social Media, Spam Detection, Twitter.

I. INTRODUCTION

An Online Social Network i.e., a Web-based application that enables users to create a public or semi-public profile inside a closed system, identify other users they are connected to, and browse and examine both their list of connections and those made by other users. Some of the Online Social Networks (OSNs) that are widely popular right now are Facebook, WhatsApp, Twitter, Instagram etc. Along with the growth of the social networks, increased the number of spammers. Spammers are the users who manipulate the platforms to broad cast unwanted or malicious messages. Twitter is a microblogging service where users can

* Authors for Correspondence

post 280-character messages called tweets. The Success of social networking services can be seen in the dominance of today's society with Twitter having 330 million monthly active users by 2020.

As of May 2020, every second, on the average, around 6,000 tweets per second or 350,000 tweets sent per minute or 500 million tweets sent every day or, 200 billion tweets per an-num are present facts. thanks to this huge growing trend, this Online Social Network has attracted many users along -side spammers. Web Attacks that have appeared on Twitter are Scam, Spam, Phishing etc., Spam may be a sort of Platform Manipulation.

Platform Manipulation is taken into account as an activity that's intended to negatively impact the people's experience on Twitter. This includes unsolicited or repeated actions. Spam can include malicious automation and other sorts of platform manipulation like fake accounts.

Shortened URL is included in most of the Spam Tweets to trick users into clicking on it. Additionally, in an effort to reach a wider audience, they frequently tweet about related trends since resources, such as tweets can be shared with each other. This type of Web At-tacks not only disturbs the user experience but also causes a whole internet damage which may possibly cause temporary of Internet Services all over the Globe.

To deal with the consequences, user can report a spam by clicking their home page. Then Accordingly the spam accounts are suspended. However, as the Total number of Tweets sent per Day are 500 million in 2020, Among which 10%(Approx.) are of Spam Tweets. This has become a major problem on finding an appropriate Solution.

Resul Kara et el [1]., They believed that in order to guarantee a spam-free atmosphere, it is necessary to identify and filter the tweets of spammers in addition to their owners. Reducing false positive detections is essential in order to prevent innocent users from being labeled as spammers. They employed a mixed classification strategy with SVM, Decision Tree, and Naïve Bayes classifiers. Additionally, Twitter's antiquated features—which are frequently used by Twitter spam detection techniques—are emphasized. Presented are a few new Twitter features that, to the best of our knowledge, haven't been covered by the other works.

Rohini et el [2]., in their paper titled Improving Spam Detection on Online Social Media with hybrid classification techniques on Twitter platform, tried to use the Naïve Bayes theorem classifier and build a speaker organization to exclude spam and not spam. In this paper they opined that Using ML algorithm SVM (Support vector machine) and NB are used to Improving Spam Detection on Online Social Media with hybrid classification techniques on the Twitter platform. The System offers a basic assessment of ML algorithms for the identification of streaming spam tweets in this dissertation. The system is used in this evaluation to process both real-time and offline tweets that are updated in real-time. The system found that one crucial step before ML-based spam detection was feature discretization.

In this Project, the one way of solving this problem is given. We have used the approach of Machine Learning Algorithms. Classification is used here; In predictive modeling,

classification is the process of predicting a class label for a given example of input data. Here we have used it to get whether a tweet is spam or not. We used an open-source dataset from Kaggle (size of 14899, 7) which contains => 7454 of “Quality” tweets and 7443 of “Spam” tweets. Initially we created a training dataset that contains information about the tweets, including some features required like following, followers, actions, is retweet, location and specific labels i.e., spam, quality to train the models and then use the trained models to detect the real-world tweets as either Spam or Quality.

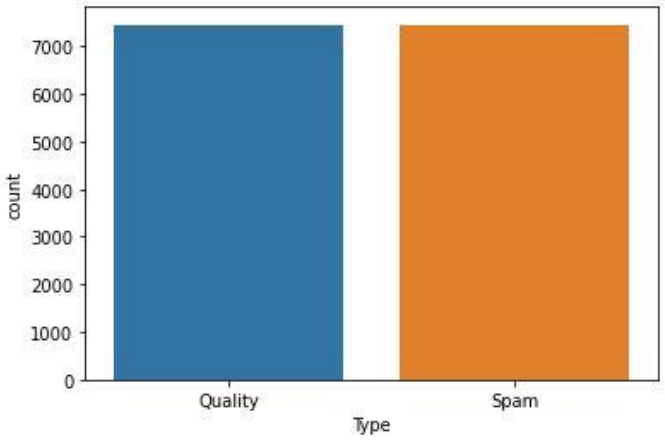

Fig. 1: Count plot of dataset showing distribution of two classes namely, Quality and Spam.

II. RELATED WORKS

There has been considerable work in the field of detecting spam for different social net-work data. In the recent past this topic was able to draw considerable attention from the researchers. Due to extensive study and expansion of machine learning based processes, classical natural language processing-based studies were done a little less, but still there are good number of studies who go for semantic based studies [1]. In various studies, it was found that behavior features are effective while detecting spams. Tang et el. applied generative adversarial networks to solve this problem [2]. Machine learning and deep learning algorithms have extensively been used to achieve higher accuracies in detecting spam messages [3] [4] [5]. Noekhah et el. proposed Multi-iterative Graph-based opinion Spam Detection (MGSD) which was found to increase the accuracy of the machine learning algorithms significantly [6].

Sentiment analysis deals with analysis of users’ views, assessments and imitations about objects, individuals, events, issues, and facilities. Sentiment analysis uses textual data collected from social networks, analyses them using various tools and techniques like natural language processing, machine learning, deep learning, soft computing to effectively identify the sentiments involved [7] [8] [9] [10]. Sentiment analysis can be used to get an idea about the ideas articulated in different posts. Sentiment analysis also categorizes the collected data

into different categories. There can be simply two types of classes, like positive and negative only, or there can also be more than two types of classes. When there are only two classes involved, it is called binary classification, on the other hand if there are more than two classes involved, then it is called multiclass classification. These types of study can be used to understand the users' opinion about various products, services, events [11] [12] [13] [14]. Though English is one of the most used languages in these social networking sites, users often prefer to use their local language to express their views and opinions. These multilingual data make the processing and analysis of the data more complex and challenging [15] [16].

III. MATERIALS AND METHODS

In this study, we propose machine learning classification techniques on Twitter data, with the focus being on detection of spam tweets in twitter. The issues are: Curse of Dimensionality, accuracy and precision. Here to face the issue of Curse of Dimensionality we have proposed "Principal Component Analysis" (PCA) to reduce the dimensions without information loss and for the classification process, we performed comparative analysis on the classification algorithms of supervised learning techniques such as XGBoost, AdaBoost, Random Forest, Decision Tree and Logistic Regression to classify the tweets into labels i.e., Spam, Quality.

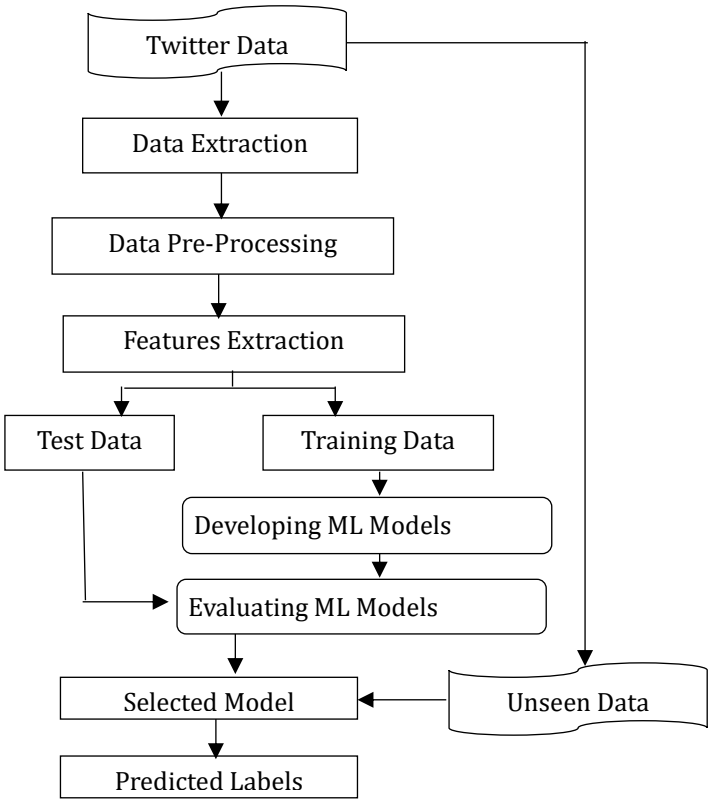

Fig. 2. The workflow diagram

2.1 Procedure for proposed method

In this study, machine learning algorithm techniques along with PCA were proposed for getting the final predictions of tweet from Twitter which helps in detection of spam tweets. The process is subdivided into 4 modules: (i) Data Collection, (ii) data pre-processing (iii) model building and evaluation (iv) Comparison of the Models (v) Prediction on Real-Time Tweets. Here Developer Twitter account is used for the extraction of the Real-Time Tweets with the help of Twitter search API using "rtweet" package in R.

2.2 Data Collection module

In this study to train our model we have collected the dataset from Kaggle[3] in which it contains all the tweet information extracted from twitter that includes [Tweet, following, followers, actions, isretweet, location and Type with specific labels i.e., spam, quality].

2.3 Data pre-processing module

a. Text Pre-Processing

Usually the tweets contain many special symbols like hashtags(#), underscores(_), URLs, @, etc. In this step; for the removal of these special symbols we used Natural Language Toolkit(NLTK) that contains the process of :

- Converting Text to Lowercase
- Punctuation removal
- White spaces removal
- Tokenization
- Remove stop words
- Stemming
- Lemmatization
- Part of speech tagging (POS)
- Build Corpus

To convert the cleaned tweets to numerical data we utilized OneHotEncoder from Tensorflow which assigns the numerical identification for each word in a tweet.

b. Principal Component Analysis (PCA)

But after performing the previous step we have faced a backlash i.e.; curse of dimensionality of encoded tweets so in order to overcome this issue PCA (Principal Component Analysis) is Introduced as dimensional reductional tool to reduce the embedded tweets data into two columns without data loss.

IV. RESULTS ANALYSIS

To Build and evaluate our models, we have split the dataset into testing and training sub-sets of data according to the percentages 33% and 67% respectively. Then these data sets were used to train the following machine learning algorithms.

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- ADABoost

The following performance measures were used to find compare the developed models

Accuracy is to inspect the accuracy of new data that has been for the trained model.

$$\text{Accuracy} = \frac{\text{True Negatives} + \text{True Positives}}{\text{Flase Positve} + \text{ True Negative} + \text{True Positive} + \text{Flase Negative}}$$

Precision is one of the standard metrics that is a measure of classifier's exactness. The lower precision denotes that it deals with a huge numerical data of false positives in the result. Precision is calculated by considering the ratio of number of True Positives and the total number of true positives and false positives.

$$\text{Precision} = \frac{\text{True Positves}}{\text{Flase postives} + \text{True Positives}}$$

Recall it is the measure of our model correctly identifying True Positives. Simply it measures the classifier's completeness. A lower recall represents presence of many false negatives in our predicted result. Its is the ratio of total true positives to that of true posi-tives and false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{False Negatives} + \text{True Positives}}$$

F1 Score is a metric used when both precision and recall metric are required to measure the performance of the classifiers. This metric measures the association between recall and precision.

$$\text{F1 Score} = \frac{2 * \text{Recall} * \text{ Precision}}{\text{Recall} + \text{Precision}}$$

The results obtained after evaluating the models are presented in the table below:

Table 1. Results

Classifier	Accuracy	Precision	Recall	F1 Score
Random Forest	0.990	0.992	0.987	0.989
Decision Tree	0.985	0.985	0.988	0.987
AdaBoost	0.994	0.994	0.985	0.990
XGBoost	0.997	0.997	0.987	0.992
Logistic Regression	0.540	0.540	0.540	0.540

V. CONCLUSION

In this paper an attempt was made to detect spam messages in twitter data using machine learning algorithms. An open-source data set was used to develop the models and some real time data were extracted from twitter to check the models developed. Major classification algorithms were used in this study and the results shows apart from logistic regression; the other four algorithms have almost similar results. The differences among the accuracy measures are very less to be considered. But if we see to the minute difference, XGBoost is found the clear winner in the league. The results obtained are totally based on the open source data set, and these results might be validated using a newly collected dataset. Also in the further studies, deep learning models can be used and results can be verified.

REFERENCES

1. N. Saidani, K. Adi and . M. S. Allili, "A semantic-based classification approach for an enhanced spam detection," Computers & Security, vol. 94, pp. 1-12, 2020.
2. Tang, , T. Qian and Z. You, "Generating behavior features for cold-start spam review detection with adversarial learning," Information Sciences, vol. 526, pp. 275-288, 2020.
3. K. Dedeturk and B. Akay , "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," Applied Soft Computing Journal, vol. 91, pp. 1-18, 2020.
4. G. Dada , J. S. Bassi, H. Chiroma, S. M. Abdulhamid , A. O. Adebayo and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, pp. 1-23, 2019.
5. Makkar and N. Kumar, "An efficient deep learning-based scheme for web spam detection in IoT environment," Future Generation Computer Systems, vol. 108, p. 467–487, 2020.
6. S. Noekhah, N. b. Salim and N. H. Zakaria, "Opinion spam detection: Using multi-iterative graph-based model," Information Processing and Management, vol. 57, pp. 1-21, 2020.
7. K. Suppala and N. Rao, "Sentiment Analysis Using Naïve Bayes Classifier," International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. 264-269, 2019.
8. P. S. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha and A. Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms," International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. 1225-1228, 2019.
9. P. P. Rokade and A. K. D, "Business intelligence analytics using sentiment analysis-a survey," International Journal of Electrical and Computer Engineering (IJECE), pp. 613-620, 2019.

10. R. Kumar and K. R. Rao, "Sentiment Analysis using Social and Topic Context for Suicide Prediction," (IJACSA) International Journal of Advanced Computer Science and Applications,, pp. 388-396, 2021.
11. S. D. P. Devisetty, Y. M. Sai, V. A. Yadav and P. Vidyullatha, "Sentiment Analysis of Tweets using Rapid Miner Tool," International Journal of Innovative Technology and Explor-ing Engineering (IJITEE), pp. 1410-1414, 2019.
12. J. S. Babu, C. S. K. Rao, D. Banerjee, S. S. Imambi and G. K. Mohan, "Opinion Mining for Drug Reviews," International Journal of Innovative Technology and Exploring Engineering (IJITEE), pp. 1314 - 1318, 2019.
13. M. Sirajuddin, S. S. Babu, R. A. L. Busi and K. . R. Sekhar, "An Effective Approach of Sentiment Analysis for Price Prediction," International Journal of Advanced Science and Technology, pp. 2268-2276, 2020.
14. S. Gogulamudi, V. Mahalakshmi and I. Sreeram, "To Improve the Efficiency in Sentiment Enlists," International Journal of Control and Automation, pp. 324 - 331, 2020.
15. K. VaraPrasad, B. S. Prasad, P. Chandrasekhar and R. K. Tenali, "Multilingual Sentimental Analysis By Predicting Social Emotions Via Text Summarization," International Journal of Recent Technology and Engineering (IJRTE), pp. 1522-1526, 2019.
16. D. Londhe, A. Kumari and E. M., "Language Identification for Multilingual Sentiment Examination," International Journal of Recent Technology and Engineering (IJRTE), pp. 3571 - 3576, 2019.