A CASE STUDY ON SPAM DETECTION FOR TWITTER COMMENTS USING MACHINE LEARNING

K. Veera Kishore[1], Vasireddy Saketh[2], Chakilam Sai Deekshith[3], Vantaku Venkatesh[4], Sriramoju Sai Prasad[5]

[1]Associate professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2,3,4,5] IV[th] Btech Student, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

ABSTRACT

Now a days, social media platforms have become an important part of our existence. The social media networks like Facebook, Instagram, Twitter, SnapChat and YouTube are used for communication among people and source of promoting businesses. Twitter is an excessive communication and sharing platform, where people can share their emotions and promote their businesses by using 140character messages. More than 42millions Twitter accounts are created every month. Twitter's receptiveness to spamming has prompted the prominence of activities on Twitter. Twitter spam is a very a sophisticated issue however it's difficult to unravel. So far, previous research has suggested a number of detection and defense methods that secure the Twitter users from spammers. So, we are going to work on spam detection techniques of Twitter. This study consists of 3 sections: 1- Background about spam detection on Twitter. 2- A literature review comparative analysis of machine learning, deep learning and hybrid algorithms. 3- Discussion on limitation of previous studies and future directions.

Keywords— Social networks, Twitter Spam detection, Feature extraction, security, Data.

1.INTRODUCTION

In recent years, millions of Internet users have been able to communicate and collaborate on social media online networks (OSN) . Today, we have entered the age of online social networks OSN. Interest in this issue has been growing misinformation spread online on social media. Facebook, Twitter, and LinkedIn are the most prominent social media platforms that enable users to communicate with each other, use information and communicate in a meaningful way. Twitter is a great communication platform and sharing, it attracts profiles when provided services for spreading 140-character messages. Every month, the number of new accounts increasing more then 42 millions on Twitter . Companies and individuals impressed the supremacy of quickly sharing information, it also performance as a smart power for the sender unsolicited and uncontroversial messages over the Twitter. This kind of data or messages are understood as spam data or messages. Though, due to the immense fame of Twitter, it also attracts the attention of cybercriminals (such as spammers). Manual filtering of messages or data from Twitter is the starting point of spam detection, then there are some popular features that can detect a spam message with the help of modest filtering guidelines. Traditional machine learning methods used for spam detection models and automated spam detection methods are also started with the utilization. For the spam filtering simple blacklisting, content-based and conversational spam detection techniques of data mining methods used. These type of methods done fairly on large data or emails messages, but, identifying the spam is being a big challenge from small and noisy spam detection platforms day by day. In short, from the domain of Twitter and SMS, it is more difficult to identify the cause of the spam, noise and small length of messages and emails. In this research, use different machine learning, deep learning techniques and compare their performance on larger datasets. Also, squeeze and compare performance as well as the number of features extracted.

2 .BACKGROUND

Social Media platforms are digital-base innovation that encouragh the sharing of thoughts, considerations and data through the structure of virtual organizations and networks. By plan, Social media platforms are we based and provides customers brisk electronic correspondence of substance. Social media platforms are an aggregate term for sites and presentation which center on correspondence, local area based information, communication, content-sharing and corporation. Social media platforms without a doubt has become a necessary .

The workplace of interchanges and showcasing manages the fundamental like Instagram, Youtube, Facebook, Twitter and SnapChat accounts. Span can be characterized as superfluous or spontaneous messages sent over the internet. These are normally sent to an enormous number of customers for an assortment of utilization cases, for example promoting, phishing, spreading malware and so forth. Spammers the entire heart has gone to this stage and versatile organizations, brining about a multiplier increase in the measure of spam. From fruad accounts deluding posts, social media spam makes superfluous clamor that overwhelms real content and commitments. Twitter is an American microblogging and person to person communication administration on which clients post and associate with massages know as "tweets". Over the previous years, Twitter has pulled in an ever increasing number of clients to post messages, turning into another style of web administrations for online correspondence and spread data. Starting at 2018, Twitter had more then 321 million month to month dynamic users. Twitter is very microblogging organization, given that most Twitter's post are composed by a minority of users. The fame of Twitter engage the spammers which have prompted to the increasing of spam. There is a number of misrepresentation or utilization of fraud accounts by spammers and advertisers. While I have describe social media spamming is undeniably more viable then traditional techniques like email spamming. Currently, fraud reviews and spam has expending and are tuning into a major issue. As per to research, 15% of the Twitter users are robot and normally one out of 20 tweets is spam. Spam on Twitter effect both online social experience and cyberspace. In September 2014, the New Zealand Internet liquefied down because of the spread of malware download spam. Such spam tempted users to tab on URLs that professed to contain Hollywood star photograph, yet actually users were told to download malware to dispatch Ddos assaults. The bellow tweet is an illustration of spam: "RT@Stormzy1: The clean hearted always win in d end. U bad mind lil weirdos wid u r bad energies are gonna destroy urselves trust", additional illustration, "Aft I finish my lunch then igo str down lor. Ard 3 smth lor. U finish ur lun`ch already?" Several techniques have been suggested to combat spam. To automatically detect spam, researchers have implemented data mining algorithms to make spam detection a classification issue. Data mining has many types but in this research, spamming extraction from tweets by using deep learning and machine learning. Machine learning uses two main techniques: Supervised learning allows you to collect.

3.LITERATURE REVIEW

In previous research, data mining algorithms are applied on tweets dataset. We have examined previous work on the bases of machine learning, deep learning and hybrid algorithms. These algorithms comparison structure is below in Fig 3.1:

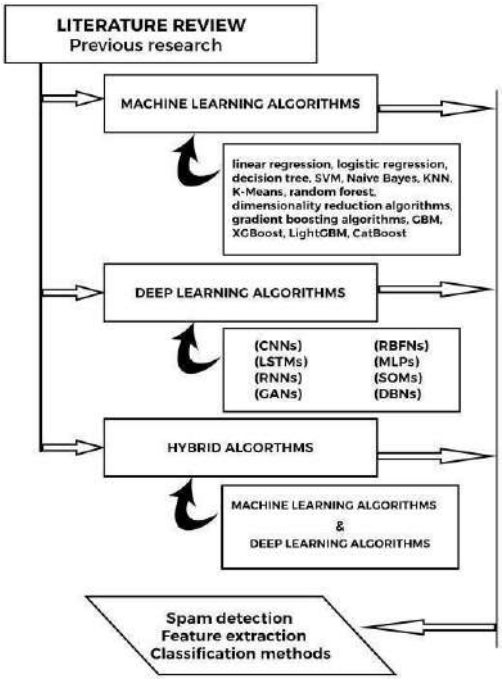

Figure 3.1

3.1 MACHINE LEARNING ALGORITHM

Proposed included extraction steps and preprocessing methods for distinguished weather tweets were spam or not spam. The feature extraction was ordered into different

detection and fake content based detection.

Fake user based detection is also compared with methods depend on a few features such five distinct classes of account information based features, user profile based feature, user interaction based feature, and user activity based feature, tweet content based features and 28 different features included. Learning process through two polynomial kernels and gaussians a support vector .

The acquire result shows the excellence of the research method by using polynomial kernels and SVM algorithms with .96 accuracy, .93 efficiency, .988 precision and F- .969. Suggested a better way to abolish misused technologies and search new ways to give results in progress. They proposed four modules: Data Evaluation that analyzes data, Pre-handling that handles the missing data in datasets, feature engineering that discounted the selection feature to machine learning algorithm and prediction module just tested the all processing step that applied on datasets not used for

training. The given architecture just tells the way of detecting spam data. They did not implement any method on the explained module, they just suggested how to detect spam data. Presented the whole process as dependent on Learning and Classifying. It categorized the Twitter spam detection approaches and afterward sorted spam tweets as URL based spam

as time features, content features, structure features and user features. The datasets about breast cancers cells that were collected from Twitter. Two classified modules applied on datasets that were SVM (Support Vector Machine) and Naive Bayes. Both comparison performance results were SVM Accuracy 83% and Naive Bayes Accuracy 92%. Hence, Naive Bayes Accuracy was higher than SVM. Introduce a new camping detection model that depends on vector-based qualities for sentence installing. The whole research depends on 3 basic steps: Firstly, to analyze the similarity of Twitter accounts in which posts or tweets are on the same topic. This similarity helps to build a graph. Second step, to classify campaigns, the graph was built on the basis of similar accounts. Third step, classifying the detecting tweets as spam campaigns. Ground-Truth Twitter dataset from Twitter obtained by using a real-life 3-day. Two-step semantic similarity function applied on datasets.

The Sent2vec model is used for found similarity and manhattan lstm model is used for recalculating the similarity. These models provided the result of 58 candidate campaigns: A Precision was 0.945, A Recall was 0.93 and AF was 0.946. These models were compared with the U & T Based Model that provided the precision was 0.909, A Recall was 0.873 and AF was 0.89. Slove the issue of categorized news articles identified with disinformation and standard news by surveying dissemination contrivances on Twitter. Italian dataset was collected from US mainstream articles and disinformation articles, IT mainstream articles and disinformation articles. Multi-layer diffusion network global networks may be viably misused to recognize online disinformation. off-the-shelf classifiers for example, logistic regression on dataset relating to two diverse media scenes (US & Italy) produce exceptionally exact arrangement results (AUROC up to 94%) which are much better than our baseline with upgrades up to 20%. Applied 5 different feature extraction on 2 different datasets, the first dataset is collected from SHP and the other one is custom collected. Feature extractions are: account based features are used to collect outer information about accounts. Stylistic features are used to identify the symmetric variations of NL. Hashtag based features allow the user to apply tagging facilitates. Word embedding based features where words have the same meaning and representation. Topic word based feature used as important keywords. The proposed model has a total 4 steps: tweets extracted from different Twitter accounts, preprocessing techniques stop words and tokenization applied on extracted tweets, Feature extraction using LSA, LDA and glove applied on collected datasets and.

in the last step datasets is ready for train test splitting. For best results applied evaluation metrics and MLP recorded the highest accuracy 93%, 98% accuracy was observed for the SPD dataset and Classification probability 97 accounts correctly classified and only 3 misclassified.

Proposed an Ontology-Based framework for criminal intention classification (OFCIC) framework for detection of spam and suspicious posts or tweets from Twitter.

Ontology of Criminal Expressions (OntoCexp) presented for execution of above framework. This research had two parts: function and content. Function part is used in OFCIC for characterized the intention of the speaker and specify the illocution. OntoCexp used a content part which presented the meaning of the post. ML techniques are used to automatically illocution class to tweets posts. The best ML configurations presented F1-score around 0.5 and the result obtained 0.72 of general F1- Score by combining glove and ANN techniques. Used a Denstream known as density-based grouping technique for sorting floods of tests. Summarized the whole model into five main steps: 1) by arriving the primary window of data, two or three bunches are made by Euclidian distance, since no genuine microcluster has been made now. 2) By arriving the second part of data, if the general population of all the made microcluster outperformace "MinC" an INB classifier will be given out to it 3) for every micro cluster whose populace surpasses "MinC," a full INB is prepared. 4) To try not to occupy the memory, the examples are killed aside from its markers like population, timestamps, and mean. 5) Updated to force an extremely low-computational complexity to the proposed system in connection with standard DenStream. Randomly four types of employee datasets collected from Twitter. SimThreshold esteems bigger than 0.8and lower than 0.5. The parameters are set within the range of [0.6–0.8]. The given methods gives the equivalent or more noteworthy outcomes then the DenStream. Proposed a hybrid approach for identifying the spam based profiles on the bases of similarity. Cluster approaches are used for selecting the initial spam accounts for classification purposes.

Three classifiers were used in the proposed model: multilayer perceptron (MLP) used to solve the linear and nonlinear classification problems, support vector machine (SVM) analyzed the data and detect the pattern, Random Forest is the branch of decision tree and it works on tree structure.

3.2 DEEP LEARNING ALGORITHMS

CNN was exploited by two classifiers. Text-based classifiers embedding the text before sending the CNN, CNN made neurons with learnable weights and biases and soft max function used for class score prediction in classification. Combined classifiers use meta-data as input, normalized input data in 0, and 1 form, combine the metadata classifier and text- tweets and send then as input for classification. Social honeypot dataset and 1 KS to 10 KN dataset were used in this study. Combined classifier provided the high accuracy rate was 99.68% and 93.12% for dataset I, II. Proposed a new architecture model with help of other three different architecture models. Firstly, Convolutional Neural Networks with semantic layers and known as Semantic Convolutional Neural Networks. By using Concept Net knowledge-based and

WordNet the initial text was enhanced and is signified with word2vec based. Secondly, LSTM neural networks with semantic layer known this framework Semantic Long Short Term Memory. It enhances the semantic representation of the words. Finally, present the combination of above two models that. model is named Sequential Stacked CNN-LSTM Model. Above models used for spam detection from social media and it take the advantages from above both models.

Twitter dataset and SMS datasets were collected for implementation of hybrid models and this model compared with traditional models and get good results. SMS dataset accuracy rate was 1.16% and Twitter dataset accuracy rate was 2.05% that was increased rate.

For detecting spams from Twitter proposed a Neural Network-based technique with traditional features-based method and deep learning methods. CNN was used for experiments with multi word embedding. Machine learning algorithms are frequently one sided toward majority class. 1KS10KN and HSpam data sets are used in this research paper.In CNN, 1KS10KN recall was low and HSpam recall was high.

The feature-based methods perform ineffectively when analyzed 14 datasets of HSpam against the deep learning methods. F- mesure was 0.984 get from results. To find the varieties of spam activities proposed a novel technique based on deep learning techniques. Word vector training mod was learned the structure of each tweet. After this on representing a dataset binary classifier based builted. In investigations, 10-day real. tweet. datasets are collected and implemented to evaluate proposed methods.

3.3. HYBRID ALGORTHMS

By using community-based feature for detection of automated spammers proposed a hybrid approach besides further features like content-, metadata-, and interaction- based features. Followings, followers and other activities of the user provide the information. The research revolves around such characterization of the spammer that is based upon its neighboring nodes and their respective interactions. For spam detection analyzed to be the most effective features were Community-based features and metadata-based but metadata is the least effective for spam detection.

They hybrid systems based on social honeypots used to detect the spam tweets, content filtering to detect similar tweets and classify the results that were provided by above two layers. The API streaming dataset of 100000 Twitter profiles that had malicious and legitimate profiles was trained by the preprocessing technique of spam filtering, text-based spam filtering, content filtering, extract characteristics and word N-gram. The model was tested by four algorithms that were random forest, bayes naive, treesJ48, classification via the regression and CNN-LSTM. Accuracy of classification regression was 99%, a positive rate & negative rate was (100%), recall and f-measure was 99%. Precision was 99% and error rate was 1.7965%. [36] Giving a spam detection system which detects spam tweets in near real time by using raw data capture. To design a training model on a large number of detecting spam tweets data for experiments. After preprocessing, real-time pulling data is used to collect 200 tweets at a time and it also helps the user to detect whether the tweet is spam or not. Before applying the above techniques, lightweight feature extraction extract 13 features on collected dataset of ground truth data. Nine machine learning algorithms used for spam or non-spam tweets and for training used ground truth data. Supervised machine learning algorithms classifications are: K-Nearest Neighbor-based algorithm, boosting algorithm Naïve Bayes, Neural Network, Deep learning, Gradient Boosting machine, Boosted Regression, Random Forests and Decision Tree- based algorithem. The probability of spam tweets combined with nine algorithms results that showed accuracy was 80% and F-measure and TPR values were above 80%. Said some researchers and industries use different approaches that base on only tweets-based features. In this research proposed a new framework that contains tweet-based features and user-based features alongside text based features for classification of tweets.

4. DISCUSSION

In previous work, more variables needed to add in the framework to enhance the accuracy of the model and classification rate. Need to improve text similarity for extracted new strange words from tweets. In previous researches, data mining algorithms were applied on small amounts of collected dataset and limited tweets. So, large amounts of data set need to be tested for the accuracy of previous algorithms. In Future, we can collect the dataset of tweets in different languages. We can apply data mining algorithms on other social media platforms like Facebook, Instagram, LinkedIn, YouTube and WhatsApp. More classifiers can be added that can make Twitter spam detection more valuable for users. Research will help to solve model scalability without performing comparative accuracy. Can use the characteristics of spammers at different levels of granularity have been used by some interesting patterns released by spammers.

5. SIGNIFICANT

Implementing spam detection is essential for any social media platform especially Twitter. Spam detection not only helps keep detect spams from tweets, but also helps . improve the quality of life of social media accounts because they run smoothly and are only used for their intended purpose. Therefore, we are going to implement data mining algorithms for detecting spam tweets, messages and URLs from Twitter.

6. CONCLUSION

Today is the time of social media and Twitter is the most well known social media network where anyone can post their thoughts, send their messages and promote their business. Followers have been increased on Twitter to capture attention of the spammers. In previous research, there are many algorithms of data mining that are used for spam detection on Twitter's collected datasets. In literature review, we have compared the different data mining algorithms in the category of machine learning, deep learning and hybrid algorithms. All of these algorithms researchers use for different types of spam detection. But the previous algorithms are not enough to extract and detect the spam on Twitter accurately. So, we need to expand the research for the high classification rate of spam detection. In future, we will apply previous methods on further social media .

7. REFERENCE

[1] Rutuja Katpatal, Aparna Junnarkar, "An Efficient Approach of Spam Detection in Twitter" 2018.

[2] XIANCHAO ZHANG, ZHAOXING LI, SHAOPING ZHU, WENXIN LIANG,"Detecting Spam and Promoting Campaigns in Twitter "2016.

[3] Olubodunde Stephen Agboola, "Spam Detection Using Machine Learn-ing" 2020.

[4] Surendra Sedhai and Aixin Sun, "Semi-Supervised Spam Detection in Twitter Stream" 2017.

[5] Claudia Meda, Edoardo Ragusa, Christian Gianoglio, Rodolfo Zunino, Augusto Ottaviano, "Spam Detection of Twitter Traffic: A Framework based on Random Forests and non-uniform feature sampling" = 2016.

[6] Xianchao Zhang, Shaoping Zhu, Wenxin Liang, "Detecting Spam and Promoting Campaigns in the Twitter Social Network" 2012.

[7] Aryo Pinandito, Rizal Setya Perdana, Mochamad Chandra Saputra, Hanifah Muslimah Az-zahra, "Spam Detection Framework for Android Twitter Application Using Naïve Bayes and K-Nearest Neighbor Classifiers" 2017.

[8] Buket Ersahin, Özlem Aktas, Deniz Kılınç, Ceyhun Akyol "Twitter Fake Account Detection" 2017.

[9] Rutuja Katpatal, Aparna Junnarkar, "Spam Detection Techniques for Twitter" 2018.

[10] Abdullah Talha Kabakus, Resul Ka