International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

A Machine Learning Method for Spam Detection in Twitter using Naive Bayes and ERF Algorithms

M. Arunkrishna, B. Mukunthan

Abstract*: In this era of machinery driven, online social media is a vast growing fact. The main social media is Instagram, Facebook and twitter. These are the media which are connecting the global as fast as other sources. It will be increase as tremendous way in future. These online social media users makes the information independently and also they can gobble the information. There are so many domains accepts the vital role of analyzing the social media. This may improves the throughput and also attain the back-and-forth competition. Now a day the people are spending their most of the time in the online social media. The vast increase in the popularity in the social media also makes the hackers to spam, thus causes the conceivable losses. The Cyber criminals are usually hack by produce the external phishing sites or the malware downloads. This became the major issues in the safety consideration of online social network and this makes the user experience as a damaged one. To combat with the issue of spams, there has been a lot of methods available, Yet, there is not a perfect effective solution for detect the Twitter spams with the exactness. In this paper , the collected tweets are classified with the help of NB and Enhanced Random Forest classifiers. The prediction is then assessed on many validation measures such as accuracy,precision and F1 score.

Keywords : Classification, ERF, Machine Learning , Spam Detection.

I. INTRODUCTION

Internet lovers use social media for useful purposes like getting useful informations getting opinions from others making friends online and to share and get inspiring new ideas. This kind of usage attracts cyber criminals to target the social media and soon it became the den for them. Until we make internet secured these cyber criminals continue to trick us. unlike traditional crimes cyber crimes are very hard to find out so more and more people (increasing number of people) victimized everyday. The best way to not be in potential victim is making the internet safe.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
M. Arunkrishna*, Research Scholar, Department of Computer Science, Jairams Arts and Science College(Affiliated to Bharathidhasan University), Karur - 639003, Tamilnadu, India. https://orcid.org/0000-0001-9310-9299 Email: arunkrishna.murugan@gmail.com
B. Mukunthan, Research Supervisor & Assistant Professor, Department of Computer Science, Jairams Arts and Science College (Affiliated to Bharathidhasan University), Karur - 639003, Tamilnadu, India. https://orcid.org/0000-0001-8452-3164 Email: dr.mukunthan.bmk@gmail.com

One of the major problem faced by social network users is spamming. The the driving force behind the use of social networks as common communication medium is the massive growth in availability and increasing use of smartphones and adaptation of 3G, 4G and Wi-Fi technologies in the country. Not only for communication, The OSNs Facebook Twitter Instagram and video sharing sites like YouTube, vimeo, Dailymotion are becoming dominant platforms for news and entertainment. The extensive use of social networks will affect the society in in both positive and negative ways so it is necessary to focuses on on the major problems for recent issues in identifying fake news or spam in social media. So we can better understand state of art methods and identify their research gaps in it.

A. Objectives

— To design and implement the better framework for the detection of twitter spam

— To improve the efficiency and strength of the framework, the optimized feature set was introduced for pre-processing.

— To improve the robustness of the framework, the considered features are extracted by the innovative feature extraction algorithm.

— To classify the spam and non-spam tweets, the novel classifier is introduced.

— To evaluate the performance of the framework, the accuracy, the TPR/FPR and the F-measure are compared with the state-of-art approaches.

B. Challenges to be overcome

— The time taken for training the datasets and the detection of spam tweets is comparatively high. To reduce the time, the suitable algorithm is introduced.

— Using more tweets for training is also complex. To rectify these issues, the more number of tweets for training is going to be considered.

— The performance of the classifier is improved by verify the optimized feature set with the Google Safe Browsing API.

II. CONNECTED WORKS

[1] Presented the innovative method to provide the better way of understanding of the spam users' behavior on Twitter. The main objective of this approach was to differentiate between the spam and non-spam social media posts.

The novelty of this proposal was to provide the feature set which are independent of historical tweets. These optimized feature set was presented for very short period of time on Twitter. Those features are related with the users of Twitter, the corresponding accounts and their pairwise engagement among each other. This work also demonstrated the efficiency and strength of the optimized feature set by compared with the general feature set for spam discovery.

[2] Improved the performance of the classifiers by provided the additional set of features to discover the twitter spammers. The Random forest (RF), Multilayer perceptron (MLP), K nearest neighbor (KNN) and Support vector machine (SVM) performances was analyzed across the very famous machine learning tools such as WEKA and RapidMiner also estimated. The experiment results on WEKA was overwhelmed than the RapidMiner for considered four algorithms. In both the cases, the RF classifier was outperformed than the other classifiers. These results are obviously helpful for the researchers in the discovery of the spam on social network.

[3] developed an approach for detection of twitter spam. By recognize the twitter spam, the approached system was provided the accurate details about the corresponding spam profiles. This system was considered the certain exclusive feature sets and also it have been verified with the Google Safe Browsing API for attain the additional security. This will improve the tweet classification performances and also detected the spams in twitter.

[4] proposed a method which was utilized the SVM method. To attain the better precision in the spam URLs detection and also attain the image spamming, the Image Spam Filtering and spam map was used. By using verbal features, host based feature and site popularity features, a URL based phishing detection system have been proposed. The used algorithms are Decision Tree, Logistic Regression, K Nearest Neighbors Classifier, Support vector machine (SVM) classifier, ANN, Random forest, bagging classifier, Gradient Boosting Classifier.

[5] presented the machine learning algorithm based on the concept of Latent Dirichlet Allocation (LDA). From this, the spam and non-spam tweets in the twitter are classified. According to the entered tweet word, by manually they have verified the 6320 number of non-spam words and 15000 number of spam words. These abilities are considered as an advantage to the taken machine learning algorithms in order to check whether the tweets are genuine or fake. Also, in this paper the various methods are employed for the detection of twitter spam that have been deliberated by evaluating the accuracy and the rate of detection.

[6] recognized the social spam in social networks by proposed a scalable spam detection system termed as an Oases. This have been achieved by using an online and scalable methods. By the two key compounds, the innovation of the proposed method was introduced. The first one was the deployment of decentralized DHT-based tree overlay for the purpose of collecting and discovering the dishonest spam from the social communities. The second innovative was, combining the spam posts properties for generating the innovative spam classifiers to vigorously separate the new spam. The Oases model was designed and implemented. The experiments have been carried out with the large-scale of real-world Twitter data. The outcomes were demonstrated the attractive load balancing, superior effectiveness, scalability in the detection of online spams for the social networks.

[7] proposed a semi-supervised framework which was named as the Spam2Vec. This model have been developed to identify spammers in Twitter. By leveraging biased random walks, this algorithmic framework have been acquired the spam illustrations of the node in the network. This spammer detection technique was significantly better on precision over the other baseline approaches.

[8] propose a model to overcome drifted Twitter spams. which means the spam which changes its properties over time. To to combat with this scenario the proposed model uses KL divergence and use MDD (Multi-scale Drift Detection) test to identify possible drifts. retraining the base class with detected results improve performance. This technique also provide better results with accuracy, f-measure and recall.

[9] Uses den-stream approach with proposed INB (Incremental Naive Bayes) classifier called as INB Den-Stream. It is a kind of stream clustering which filter the spam by categorising tweets as spam and non spam clusters. In Stream clustering methods which cluster may have number of micro clusters. The distribution of micro clusters may symmetry or asymmetric in its nature. This proposed method also replaces euclidean distance by set of classifiers. Here it is the proposed INB Den-Stream. The effectiveness of the proposed model is compared against with denstream, StreamKM++ and Clustream classifiers.

[10] proposed semi supervised learning techniques for spam detection. This semi-supervised approach for classifiers to stop the entire framework uses PDS(probabilistic data structures). Such as QF (quotient filter) and LSH (locality sensitive hashing) QF is also used for query the URL database and spam words database. The local sensitive hashing is used to to perform similarity check. The framework minimises the computation process. the resultant values from the parameters like precision, F-score, and recall proofs the model as a successful one.

[16] uses hybrid approach to detect spell in Twitter.This multi tier approach uses some key information from feature set and then analyse it. Additionally the model uses Google safebrowsing API for enhanced security purposes. This system uses twitter4j API and combination of NB naive Bayes and support vector machines SVM and also uses unique feature assets to provide desired solution.

III. METHODOLOGY

A. EDA (Exploratory Data Analysis)

EDA is a well-proven approach to perform primary enquiry on data . It employees variety of techniques and procedures to carry out data analysis with EDA we can build patterns eradicate anomalies and do away with hypothesis. So we can predict insights from the data. The assumptions are made with the help of statistics and results are in the form of graphical techniques. EDA Consisting of the methods such as plotting raw data, plotting statistics, positioning of plots to maximize pattern recognition abilities.

It is important process to understand and to relate it with business perspective. Word cloud - open source tool Steps in this section

a) get tweet /spam data from available data sets: Data sets are records that holds collection of instances. It is organized in to some type of data structure and they are related to a particular type of information.

Twitter spam detection datasets contains two kinds of data: ham and spam. Sometimes ham messages are very hard to predict its authenticity. That type of messages are known as hard ham i.e. ham but hard to know it is. so it is falsely identified as spam. UC Irvine machine learning repository,Kaggle data set, AWS data set and many others providing varieties of datasets for free .

b) Explore and analyze data: It is the most important and initial step for data analyst to understand what is in a dataset. It will bring out the main characteristics of data. So it is essential to look for meaningful patterns and characteristics from the large scattered dataset.

c) Visualisation: It gives the ability to translate everything in to visually understandable formats such as charts,plots,diagram etc, so that every one can easily understand it. In twitter data analysis ,visualization helps us estimate the words that have highest accuracy so it will help us to identify better model for development and deployment. One of the popular visualisation model is N-gram model.

N-gram model visualisation: It identifies how many number of words are considered as single unit then it split the data set into two sets a)training set and b)testing set. word cloud is one of the popular tool that simplify visualisation process.

B. Data Preprocessing

Data preprocessing in machine learning involves the action of transforming raw data into an understandable one. Noise in the data will mix it unreliable for training so we have to to eradicate noise in our data by performing preprocessing. It includes cleaning normalisation selection feature extraction and word embedding. Preprocessing helps to achieve better outcomes in ml models Text cleaning is used to remove noise from the data set such as punctuation, whitespace, numbers, hyperlinks etc. Standard procedures include converting all to lowercase, removing numbers removing punctuation and white spaces. Word streaming and word lemmatization also to be performed. Word normalization is the process of preparing text document for NLP tasks. Stemming and lemmatization are two popular normalization techniques which helps us to identify the root forms of the word.

a) Word stemming : stemming algorithms working by removing end or beginning of the words using a list of common prefix and suffix that language uses. stemming can be successful most of the time but not always because this approach has some limitations

Table- I: Word Stemming

Word	Suffix	Stem
running	-ing	run
runs	-s	run
consolidated	-ated	consolid

b)Word lemmatization: lemmatization reduce inflectional form and find the root form with the help of vocabulary and morphological analysis of word. It is done by utilising a dictionary of particular language and convert the words back to its base form. implementation of these to algorithms might be quickly because it needs lots of thinking and pre-planning but NLTK library provides implementation of these to algorithms with Ease.

Table- II: Word lemmatization

Word	Morphological Info.	Lemma
studies	Present tense	study
ran	Past tense	run

C. Feature Extraction - Preparing Text Data for Machine Learning

Text data needs special preparation. It must be passed to remove words. This technique is called tokenization. scikit-learn perform tokenization and feature extraction clear we cannot directly cook with text in machine learning so it is necessary to convert text to numbers. Well known method bag of words ECM model that concentrate on occurrences of word in a document.

Most algorithms accept input to be in integers or float so feature extraction layer converts word to ‘int’.There are popular ways to do that such as countvectorizer, Tfdifvectoriser, word embedding.

a) Countvectorizer: it changes text to word count vectors which uses dictionary of all words to relative ID and ID will relate to the word count. if suppose for the values {1:’a’, 2:’b’, 3:’c’}, and the word is ‘abbc’ then, output will be [1,2,1]. But the drawback of the countvectorizer is, it counts the common occurring words such as ‘the’,’a’,’an’ etc.

b) Tfidfvectorizer: Term Frequency Inverse Document (Tfidfvectorizer) used to overcome the drawback of countvectorizer, this algorithm can be used it simply, the words such as,’the’,’a’,’an’.

c) Word embedding: It converts words into vectors that is into vector is the format and it shows the position of the word in high dimensional space

King - man + women = queen
Delhi - India + France =Paris
Running - writing + rate =ran

The well known technique to do this is word2vec.

D. Algorithm implementation

There are varsity of algorithms available so we should do a literature review by reading multiple canonical and fixed descriptions of the algorithm. Choosing the well-suited model can make the implementation half done. Carefully choose the implementation language because it directly influence on APIs and libraries in the implementation.
Training: training gives the model ability to predict by its own. Deep learning algorithms for trained using training data set. It will create a model and predict new insights based on the trained model. Machine learning models are categorised into three types which type has its own techniques for training. Algorithm perceive patterns in the training data. So that the system can use that target attributes to map the input data with training data and get prediction on new data.

The quality of the training data must be maintained truly for successful prediction.

IV. MACHINE LEARNING APPROACH

classification in Machine Learning involves the task of grouping the observed tweets based on the set of learned values.

It works based on the training dataset. i.e. testing tweet information is matched against the training data to classify it. There are two types of classification

- Supervised classification
- Unsupervised-classification
- Semi-Supervised classification

The supervised classification uses the training data to analyze the observed image .So we need to train the system on each image type. And then we can match the testing set against the trained set.(figure1).

The unsupervised learning does not need the training set. It simply makes the cluster data for each type and assigns the class value for each cluster. Figure1 shows the supervised and semi-supervised classification, uses both labled and unlabelled data for classification. Semi-supervised classification utilise very minimal training data and based on that it self train itself.

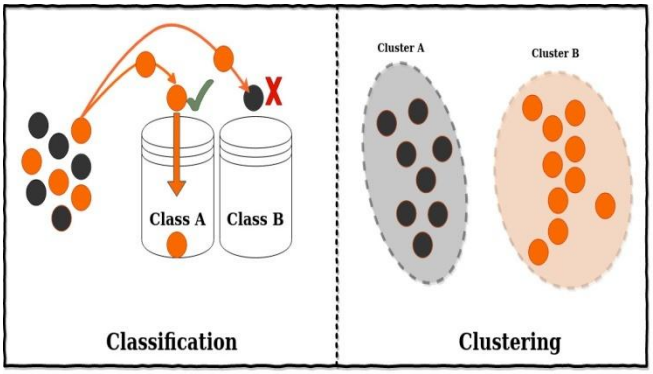

Fig. 1.Classification and Clustering

*A) Naive Bayes algorithm: N*aive Bayes classifier also known as simple Bayes or independence Bayes is an algorithm that uses Bayesian theorem to classify objects. It is a classical ML approach and has been widely used for spam filtering. It takes strong or weak (naive) independence between attributes of data points. We can use it for text classification sample detection and medical diagnosis. most of the machine learning models for based on Bayesian statistics. Naive Bayes classification is commonly used classifier which works based on machine learning. This probabilistic classifier make use of posteriori decision rule of Bayesian model. It is popular for text classification and spam detection.

With the features (x_0,x_1,…,x_m) and the classes (C_0,C_1,…,C_n), The model determines the probability of features occurring in each class. And the classifier returns most likely class.

So for each class we calculate probability distribution like $P(C_i \mid x_0,x_1,\ldots,x_m)$ for each class .So we use bayesian rule.

$$P(A \vee B) = \frac{P(B \vee A)P(A)}{P(B)}$$

here A denotes Classes ,and B denotes Features.So we can replace A with class (C_0,C_1,…,C_n), and B with the features (x_0,x_1,…,x_m). P(B) is the normalization, but it is unable to calculate.

instead, we can take,

$$P(C_1 \mid x_0, x_1, x_m) \propto P(x_0, x_1, \ldots, x_m) * P(C_i)$$

where $P(C_i)$ is a portion of dataset which falls under class i, which is easy to calculate. But, $P(x_0, x_1, \ldots, x_m \mid C_i)$ is difficult to compute. Inorder to simplify it's computation we assume that (x_0 , x_1, ... , x_m) are conditionally independent for given C_i So we can say ,

$$P(x_0, x_1, \ldots, x_m \mid C_i) = P(x_0 \mid C_i) * P(x_1 \mid C_i) \ldots P(x_m \mid C_i)$$

and it is not always true, hence the name Naive Bayes classifier.

Final representation of class probability is as follows,

$$P(C_i \mid x_0, x_1, AI) \propto P(x_0, x_1, \ldots, x_m \mid C_i) * P(C_i).$$

I.e.,

$$P(C_i \vee x_0,x_1,\ldots x_m) \propto P(C_i) \prod_{j=1}^{m} P(x_j \vee c_i)$$

So calculating $P(x_j \mid C_i)$ will depend on what distribution our features follow. In text classification it is word count. So it follow multinomial distribution. if the features are continuous then it follows Gaussian distribution.

Advantages:As compared to other algorithms naive Bayes algorithm needs very little explicit training. naive Bayes classification algorithm can able to work with high dimensional data points/ large number of data points

Classification method: the method of classification is very simple but effective. Classification of estimate probability of the given data point and comparative with classes. Fix the C_i based on the largest probability. So,

$$y=\text{argmax}P(C_i) \prod_{j=1}^{m} P(x_j \vee C_i)$$

it is referred as maximum posteriori decision rule.

Posterior probability: in Bayesian statistics posterior probability (of a random event) is the conditional probability which is calculated after that evidence for background is found and taken into account.

Maximum posterior probability:It is the estimation of unknown quality that equals the mode of posterior distribution probability

B) Enhanced Random Forest(ERF) Algorithm:

Random forest is an ensemble method/algorithm for supervised classification. it works based on decision tree classifier. This model is used to classify the instances when the class feature is unknown A decision tree is a basic building block of random-forest classifier. It can be used for both classification and regression problems to stop random forest algorithm constructs decision tree. The accuracy of the algorithm is directly proportional to the number of decision trees. In other words higher the number of decision trees, more robust result. Random forest algorithm will formulate some set of rules with the help of information gain and Gini index. The entropy is used to measure uncertainty in our data. It is known as randomness. Higher the splits, the better our prediction will be. Entropy h is measured as.

$$H = -\sum p(x)\log p(x)$$

Here, H is the entropy, p(x)is the percentage of group that belongs to a particular class.

If we have more number of classes the entropy will be high otherwise, it will be low and they are mostly depends on a single class. Entropy plays a vital role in estimating IG. IG(Information gain) is a measure that shows how much information can we attain from a class with the given feature.

$$\text{Gain(S,D)=H}(S) - \sum_{V \in D} \frac{(V)}{(S)} H(V)$$

Here S is the actual set and D is the split V is the subset of S. The best split can be identified using information gain.

Gini indicates the impurity of the north the value of Gini will be low at leaf level of the decision tree.

Gini impurity of the node is,

$$I_G(n) = 1 - \sum_{i=1}^{J} (P_i)^2$$

The Gini impurity is 1 minus the sum of J of P_i squared.CART along with random forest gives the enhancement and it is best suited for non linear relations. It can be utilised to calculate regression and classification problems. Because CART is more sensitive to the target variable than the independent variables (predictors).

Algorithm

Step 1 : *Parameter Initialization*
Folds:
Seed:
No. of trees:
Max. Depth of tree:
Br:
Step 2 : *import data set into buffer Br.*
Step 3 : *Buffer reader Br=null;*
Step 4 : *Br= buffer reader(file reader);*
Step 5 : *Extract Features.*
Step 6 : *Set no. of trees=10;*
Step 7 : *No. of folds= 10;*
Step 8 : *Set max depth=0*
Step 9 : *Evaluate results through cross validate model(data);*
Step 10: *Calculate TP, FP, F-Measure to Evaluate results.*
Step 11: *Distinguish Spam and Ham.*
Step 12: *End*

V.SYSTEM DESIGN

The better framework for the detection of twitter spam namely, A Novel Twitter Spam Detection System is proposed. Primarily, as in the [1] optimized features are selected for preprocessing. These optimized feature set was presented for very short period of time on Twitter. Those features are related with the users of Twitter, the corresponding accounts and their pairwise engagement among each other. The reason behind these feature set selection is, it have been proved as better in terms of efficiency and strength when compared with the typical feature set for spam discovery. The features are extracted by the CountVectorizer. Among the neural networks learnings, this algorithm is foremost considered and used. This is the most famous learning approach which is have the ability to hand the large learning issues. The extracted features are then classified by Enhanced Random Forest Classifier. It also have the ability to use the parallel resources as well as it is probable to attain the real time training and testing tasks. The RF classifier was outperformed than the other classifiers, in the WEKA and RapidMiner those are the very famous machine learning tools.

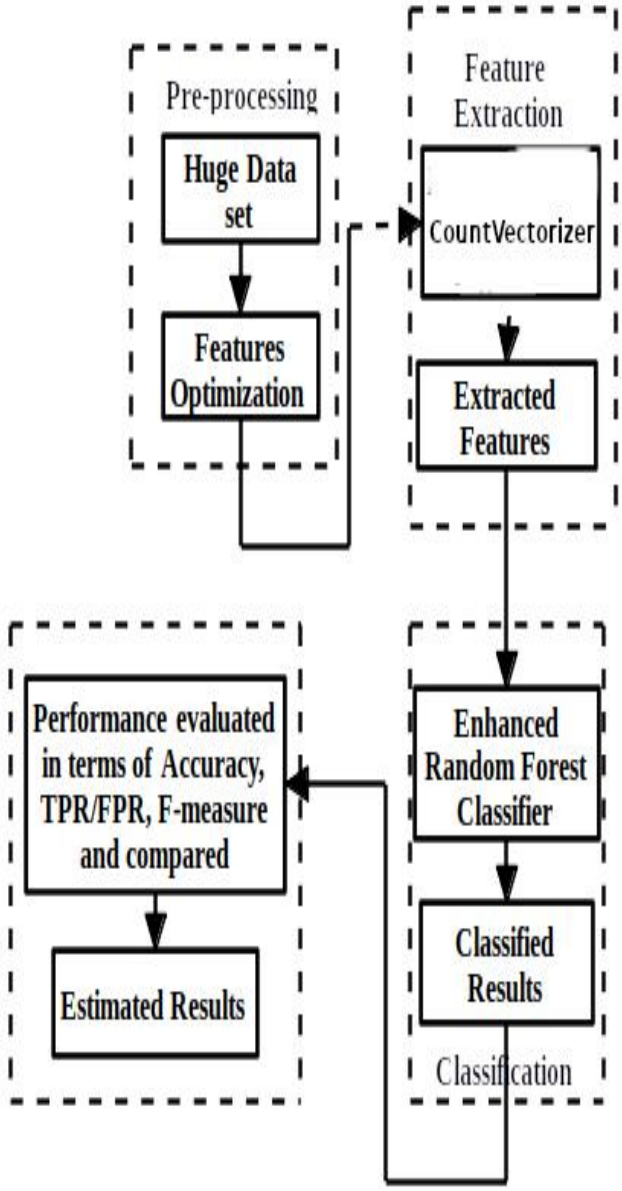

Fig. 2.Flow of proposed Spam Detection System.

VI. SCORING AND METRICS

Once training is complete, it's time to ascertain the model. Evaluation allows us to check our model against data that has never been used for training. this is often where that dataset that we put aside earlier comes into play. This evaluation metric allows us to ascertain how the model might perform against actual real world data. Accuracy alone is not the metric for the perfect evaluation. For example if a data set contains 20 spams out of hundred and our algorithm predicts all messages as non spam then the accuracy here is 80 percentage. if a data set contains 1 spam out of hundred and algorithm predicts all messages as non spam then the accuracy is 99 percentage.

So it is useless to measure performance solely based on accuracy so that we can use precision and recall

Precision: what proportion of positive identification is actually correct.

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}}$$

Recall: True positive rate(TPR) is also known as sensitivity or recall is the ratio between correctly and wrongly predicted items of true positive. i.e. what proportion of actual positive was identify correctly.

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}}$$

Confusion matrix: confusion Matrix / Error Matrix gives the summary of prediction results which are summarised by count values for each class. Centre table with four different combinations of predictive and actual values. F-measure,accuracy,recall,prediction are the important metrics on performance analysis.

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

confusion matrix is used to understand results and to visualise the system performance. so we can plot and visualise the result. Scikit-learn provides some cool plotting techniques.

$$F-\text{Measure}=\frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall+Precision}}$$

F-measure / F1 score of the system is measured as the weighted harmonic mean(average) of the precision and recall.

VII. RESULTS AND DISCUSSION

Results are carefully gathered with the help of various performance measures like accuracy, precision and f-measure. We compare the results of the algorithms with 20% and 40% training data. Our model intuitively performs well and provides above 90% accuracy on the classification process. Comparison of Precision Score. So it reveals that the proposed model achieved the best performance measures. The results of the models are given below. Here, our model achieves above 95% accuracy score and with optimal training, precision and recall values of Naive Bayes gives around 45% and ERF gives above 90 percentage. Similarly,With optimal training the F1 score of NB is around 60 percentage whereas ERF gives above 85 percentage. The results of the experiment clearly shows the improved performance of the proposed approach.

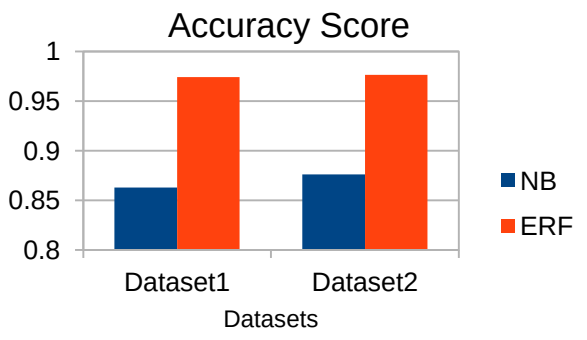

Fig. 3. Comparison of Accuracy

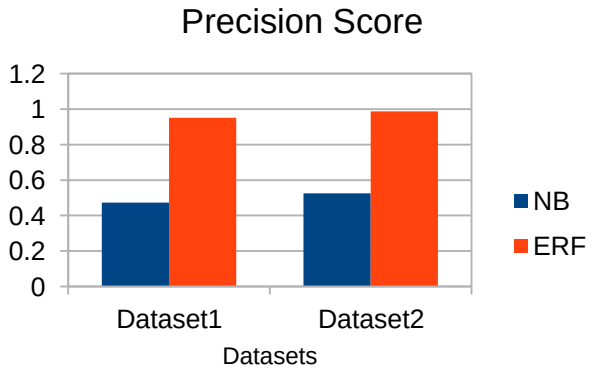

Fig. 4. Comparison of Precision Score

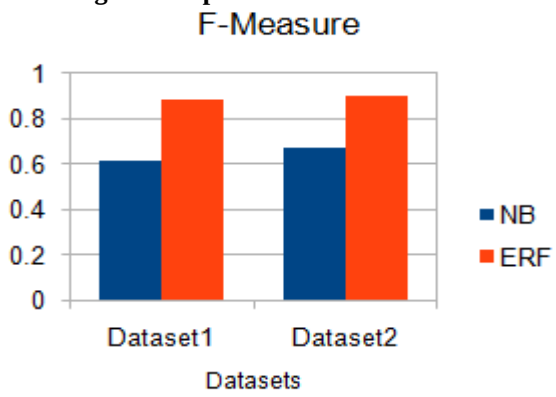

Fig. 5.Comparison of F1-Score

Table- III: Result Analysis Table

Performance Measures		Naive Bayes	ERF
Accuracy	Dataset1	0.8627802690582	0.97399103139013
	Dataset2	0.8761776581426	0.97622252131000
Precision	Dataset1	0.4724409448818	0.95081967213114
	Dataset2	0.5254237288135	0.98828125
F-Measure	Dataset1	0.6106870229007	0.88888888888888
	Dataset2	0.6690647482014	0.90518783542039

VIII. CONCLUSION

In this proposed work, we have developed a spam detection system for Twitter. For that, Machine learning approaches has been proposed. The existing methods of social spammer identification are mostly based on twitter profile data and social honeypots. There are also works using user accounts,historical tweets ans social graphs. But the increasing use of OSNs(Online Social Networks) and support for rich media messages makes it very vulnerable to spams. With the latest technologies ,spammers can easily evade or surpass traditional defense techniques. There fore more complex defenses like ML based defenses are needed. The proposed ML based framework has been explored and analyzed on basis of classification performance. To show the performance of proposed system, the F-measure, accuracy, the true/false positive rate also evaluated and compared with the existing Twitter spam detection system. In the future, we will incorporate more ML and Deep-learning methods to optimize the spam detection process

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. M. H. M. Hanif, K. S. Adewole, N. B. Anuar, and A. Kamsin, "Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner," *Advanced Science Letters*, vol. 24, pp. 1043-1046, 2018
3. V. Vishwarupe, M. Bedekar, M. Pande, and A. Hiwale, "Intelligent twitter spam detection: a hybrid approach," in *Smart Trends in Systems, Security and Sustainability*, ed: Springer, 2018, pp. 189-197.
4. P. Parekh, K. Parmar, and P. Awate, "Spam URL Detection and Image Spam Filtering using Machine Learning," *Computer Engineering*, 2018.
5. K. Madhan and K. Narayana, "A Survey of Spam Detection on Twitter Using LDA Algorithm," 2018.
6. H. Xu, L. Hu, P. Liu, Y. Xiao, W. Wang, J. Dayal, *et al.*, "Oases: An Online Scalable Spam Detection System for Social Networks," in *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, 2018, pp. 98-105.
7. S. K. Maity, S. KC, and A. Mukherjee, "Spam2Vec: Learning Biased Embeddings for Spam Detection in Twitter," in *Companion of the The Web Conference 2018 on The Web Conference 2018*, 2018, pp. 63-64.
8. X. Wang, *et al.*, "Drifted Twitter Spam Classification Using Multiscale Detection Test on KL Divergence," *IEEE Access*, vol. 7, pp. 108384-108394, 2019.
9. H. Tajalizadeh and R. Boostani, "A novel stream clustering framework for spam detection in twitter," *IEEE Transactions on Computational Social Systems*, vol. 6, pp. 525-534, 2019
10. A. Singh and S. Batra, "Ensemble based spam detection in social IoT using probabilistic data structures," *Future Generation Computer Systems*, vol. 81, pp. 359-371, 2018.
11. *Concone, F., Re, G.L., Morana, M., & Ruocco, C. (2019). Twitter Spam Account Detection by Effective Labeling. ITASEC.*
12. *Rajaraman, Anand and Jeffrey D. Ullman. "Mining of Massive Datasets." (2011).*
13. *Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space." CoRR abs/1301.3781 (2013): n. pag.*
14. *Pedregosa, Fabian et al. "Scikit-learn: Machine Learning in Python." J. Mach. Learn. Res. 12 (2011): 2825-2830.*
15. *Miller, Zachary et al. "Twitter spammer detection using data stream clustering." Inf. Sci. 260 (2014): 64-73.*
16. *V. Vishwarupe, et al., "Intelligent Twitter spam detection: a hybrid approach," in Smart Trends in Systems, Security and Sustainability, ed: Springer, 2018, pp. 189-197.*
17. *M. Şimşek, O. Yılmaz, A. H. Kahriman, and L. Sabah, "Sahte Twitter Hesaplarının Yapay Sinir Ağları ile Tespiti," Artificial Intelligence Studies, vol. 1, pp. 19-22, 2018.*
18. *K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: a tool for fake news collection, detection, and visualization," Computational and Mathematical Organization Theory, vol. 25, pp. 60-71, 2019.*

AUTHORS PROFILE

M.Arunkrishna , has received his Bachelor of Science in Information technology from *Madurai Kamaraj University* - Madurai, India in 2011 and Master of Science in Information Technology from The Gandhigram Rural Institute,(Deemed University) - Gandhigram, India in 2013 and Master of Technology in Information Technology from Kalaslingam University - Krishnankoil, India in 2015. He has 2 years of experience as an Assistant Professor in the Department of Computer Science at Valluvar College of Science and Management. His research work focuses on Data Science, Machine Learning , Deep learning . His Orchid ID is https://orcid.org/0000-0001-9310-9299.

B.Mukunthan pursed Bachelor of Science (Computer Science) from Bharathiar University, India in 2004 and Master of Computer Applications from Bharathiar University in year 2007 and Ph.D from Anna University - Chennai in 2013. He is currently working as Research Advisor in Department of Computer Science, Bharathidasan University, Tiruchirappalli since 2016. He is a member of IEEE & IEEE computer society since 2009, a life member of the MISTE since 2010. He has published more than 10 research papers in reputed international journals. He is also Microsoft Certified Solution Developer. His main research work focuses on Algorithms, Bioinformatics, Big Data Analytics, Data Mining, IOT and Neural Networks. He also invented a Novel and Efficient online Bioinformatics Tool and filed for patent. He has 12 years of teaching experience and 10 years of Research Experience. His Orchid ID is https://orcid.org/0000-0001-8452-3164.