

An Efficient Fake News Detection System Using Machine Learning

A.Lakshmanarao, Y.Swathi, T. Srinivasa Ravi Kiran

Abstract: Social media plays a major role in several things in our life. Social media helps all of us to find some important news with low price. It also provides easy access in less time. But sometimes social media gives a chance for the fast-spreading of fake news. So there is a possibility that less quality news with false information is spread through the social media. This shows a negative impact on the number of people. Sometimes it may impact society also. So, detection of fake news has vast importance. Machine learning algorithms play a vital role in fake news detection; Especially NLP (Natural Language Processing) algorithms are very useful for detecting the fake news. In this paper, we employed machine learning classifiers SVM, K-Nearest Neighbors, Decision tree, Random forest. By using these classifiers we successfully build a model to detect fake news from the given dataset. Python language was used for experiments.

Index Terms: fake news, machine learning, python.

I. INTRODUCTION

Social media is a very fast-growing thing from the last decade. Most of the information generating today come from social media. In some cases, social media can have the capability of spreading the news more quickly than newspaper Media, TV media. It can cover news that was unable to cover by other media. Generally, this kind of fake news is created to promote some agenda [8]. Fake news also caused issues like sarcastic articles or fabricated news or some cases pretending to plan government propaganda. It is very possible that 2 different articles that are similar in their number of words may be opposite in their meaning. The data science community has responded by taking action against the problem. Kaggle conducted a competition named as the "Fake News Challenge" and Facebook is using Artificial Intelligence to filter fake news stories out of users' feeds. Fake news classification is a text classification problem with a straightforward proposition. There is a challenge to design a model that can detect fake news or real news. It is very possible that 2 different articles that are similar in their number of words may be opposite in their meaning. The data science community has responded by taking action against the problem. Kaggle conducted a competition named the "Fake News Challenge" and Facebook is using Artificial Intelligence to filter fake news stories out of users' feeds.

Revised Manuscript Received on August 05, 2019

A.Lakshmanarao, Department of Computer Science & Engineering, Raghu Engineering College, Visakhapatnam, A.P, India

Y.Swathi, Department of Computer Science & Engineering, BABA Institute of Technology & Sciences, Visakhapatnam, A.P, India

Dr. T. Srinivasa Ravi Kiran, Department of Computer Science, P.B.Siddhartha College of Arts & Science Vijayawada, India

II. LITERATURE SURVEY

Misleading articles studied by Conroy, Rubin, and Chen [1]. They have shown that simple content n-grams and shallow parts of speech tagging are not sufficient for classification methods. Feng, Banerjee, and Choi [2] got more accuracy deception classification tasks. Junaed Younus Khan et.al [3] applied online N-gram analysis and machine learning techniques for the detection of fake news. Three types of fake news identified by Rubin et al. They applied various text analysis and predictive analysis methods. News is generally related to text data. So, Text data handling can be done by Machine Learning Natural Language Processing (NLP) techniques. For detecting fake news, ML NLP is a very useful tool. Hadeer Ahmed [7] used Support Vector Machines with Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction method and got good accuracy. Shlok Gilda [9] et.al applied different machine learning techniques for fake news detection.

A various number of machine learning algorithms are applied for fake news detection [5] earlier. Each algorithm has its advantages. So comparing all algorithms is necessary. Dataset is tested using Support Vector Machines, Stochastic Gradient Descent, Gradient Boosting, Bounded Decision Trees, and Random Forests. We propose a method for "fake news" detection and ways to apply it on Facebook, one of the most popular online social media platforms. This method uses the Naive Bayes classification model to predict whether a post on Facebook will be labeled as REAL or FAKE.

Raw datasets collected for fake news detection usually contain some noise such as missing values. The performance of any machine learning depends on the input data. So data preprocessing step plays a vital role before applying a machine learning algorithm. Machine Learning Data preprocessing handles missing values efficiently. Specifically, we have successfully handled the missing values problem by using data imputation for both categorical and numerical features. Selecting a dataset is an important step as the complete process is dependent on the fields, records, and data of the dataset. The dataset we used is from kaggle fake news challenge -1. The data is derived from the Emergent Dataset created by Craig Silverman. It has three fields namely headline, body text, and label. The label shows whether the news is classified as fake or real. Conroy[6] et.al constructed a fake news detection system with machine learning naïve Bayes classifiers.

A new post of facebook was tested and obtained an accuracy of 74%.

II. RESEARCH METHODOLOGY

Machine learning algorithm are using in everywhere. ML algorithms generally operates on mathematical operations. If given input dataset contains numerical values, and then applying algorithm is very easy task. But, if dataset contains categorical data, we need to apply some transformations before applying ML algorithms. As fake news detection dataset involves textual data, A special processing should be done. ML provides Natural Language Processing techniques for handling textual datasets. Analyzing the key properties of a dataset with the model best suited for that problem gives good results. Most of the machine learning algorithms require datasets to be numeric. Since the datasets in natural language processing (NLP) tasks are usually raw text, as is the case for this fake news detection problem. We used preprocessing techniques to refine the textual data, as well as a brief synopsis on the field of sentiment analysis to motivate the idea that sentiment scores can be important features for this study since they provide insight about the motivation and purpose of a piece of text. For Machine learning with fake news detection, Text documents should be represented in vectorized formats. For handling raw text machine learning provides different options. The bag-of-words technique depends on searching the whole dictionary and verifying template matching. But the problem with the bag of words method is that each language contains a large dictionary of words on file, so time taking for searching is also high. It will also fail if none of the words in the training set are included in the testing set. It requires a full dictionary for each language for the detection of the sentence. N-gram is one of the models used for text processing. The problem of language detection is that human language (words) has structure. For example, in English, it's very common for the letter 'u' to follow the letter 'q,' while this is not the case in transliterated Arabic. The n-gram model uses this notation. Particular combinations of letters are more likely in some languages than other languages. This is the basis of the n-gram classification. In the N-gram model, N can be any numeric number (1, 2, 3...). N-gram indicates that the sequence of N number of words. Assume that $n=2$ case, and there are 26 characters in the English alphabet, so possible bigrams are 26^2 (676). When compared to the bag of words technique, the N-gram method requires needs a small database. Conroy, N. J., Rubin[6] et.al finds accuracy degree in the given news and based on that they divide the news as fake or real. They used two different methods, one is a linguistic technique and the other is a network analysis technique. Selecting a dataset is an important step as the complete process is dependent on the fields, records, and data is also a supervised learning algorithm. Decision Tree classifiers are more popular because tree analysis is easy to understand.

Proposed model:

of the dataset. The dataset we used is selected from kaggle fake news challenge -1. The data is derived from the Emergent Dataset created by Craig Silverman. It has three fields namely headline, body text, and label. The label shows whether the news is classified as fake or real. There are many datasets available such as crowd sourced fake news dataset, datasets which has satire, comic labels, etc. We applied machine learning algorithms for fake news detection. We compared different classification algorithms for the detection process. We compared Support Vector Machine, Decision Tree classification, K-Nearest Neighbor, Random Forest classification algorithms. Deep Learning algorithms may also give better results for fake news detection[10], but implementation of deep learning algorithm is more complex when compared to machine learning techniques. Deep learning techniques are computationally cost effective than machine learning algorithms.

Support Vector Machine:

Support Vector Machine (SVM) can be used for regression and classification problems. In regression, SVM predicts a value, whereas in, classification, It is used to predict a class label. SVM is a supervised machine learning algorithm meaning that machine trained with training examples and later trying to predict for new test samples. In SVM, an n-dimensional space is used to plot each data item. Then, a hyper plane which differentiates the classes is used to perform a classification task.

K- Nearest Neighbor Classification:

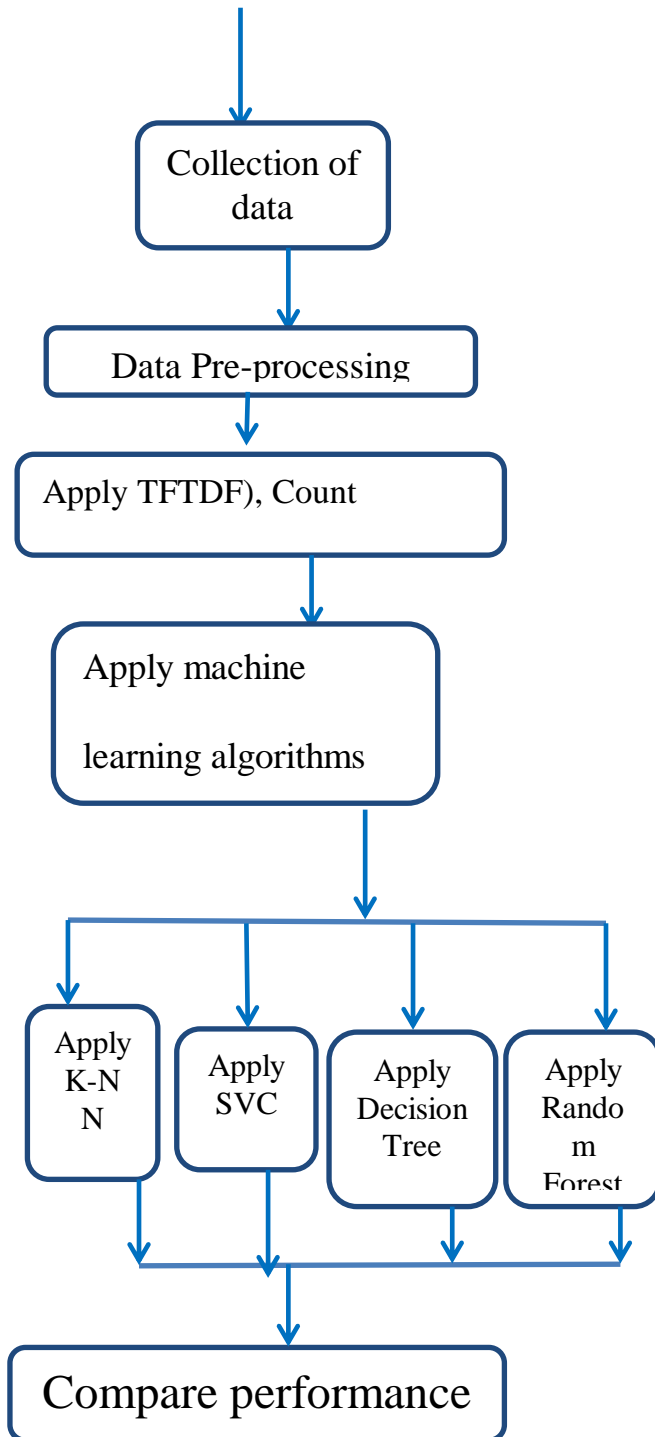
K-NN is a supervised learning classification algorithm. K-NN verifies similar things near to each other. In K-NN, K indicates number of nearest neighbors. Initially, select k value and group the data items into k groups based on similarity (distance). The items can be classified at the end. Distance can be calculated using Euclidean distance.

Decision Tree Classification:

One of the most widely used classifiers is Decision Tree Classifier. It is also a powerful classifier. Similar to SVM, Decision Tree can also perform both regression and classification. It is also a supervised learning algorithm. Decision Tree classifiers are more popular because tree analysis is easy to understand. It divides the given data set into small parts and a decision tree is incrementally constructed. The leaf nodes of a decision tree represent the classification. Decision trees are comfortable with numeric and categorical data.

Random forest classification:

One of the most widely used classifiers is Decision Tree Classifier. It is also a powerful classifier. Similar to SVM, Decision Tree can also perform both regression and classification. It

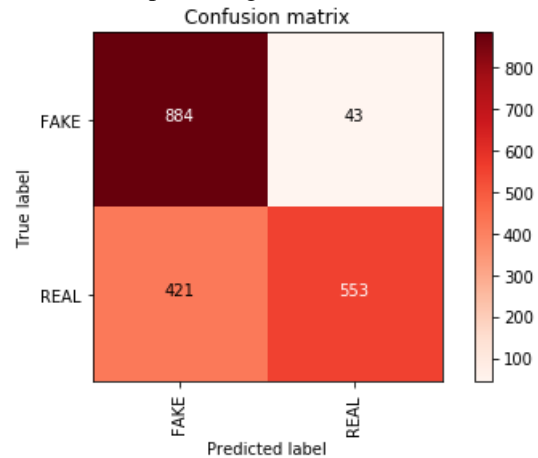


IV. EXPERIMENTS

All Experiments are conducted in python. Python was most widely used tool for machine learning. R programming is another option. Fakenews dataset is taken from kaggle. Dataset contains four features namely id, title, text, label. Dataset consists of 7796 entries. After preprocessing, 6335 rows of data is considered for analysis. Dataset is divided into training set and test set. Out of 6335 entries, 4434 entries are taken as training set and 1901 elements are considered as test set. After that, feature extraction can be done by using python NLP packages Count Vectorizer, TfidfVectorizer. In this step, Stop words also removed from the data.

Applying SVM:

Now the data is ready for analysis. First, we applied support vector machine. Performance of a classification model can be compared by a tabular format known as confusion matrix. Counts of correct and incorrect predictions are placed in confusion matrix. It shows the way in which our model is confused for predicting results.



Confusion matrix_SVM

Figure.2

Here, TP (True positives) = 884

FP (False Positives) = 43

FN (False Negatives) = 421

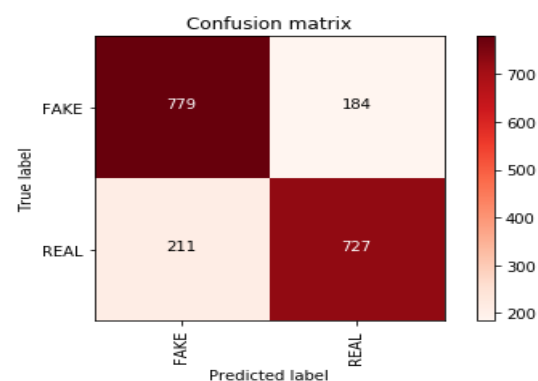
TN (True Negatives) = 553

Classifier accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{884+553}{1901} = 75.5\%$

Applying K-NN:

We applied same preprocessing steps for K-NN also.

The confusion matrix and accuracy of K-NN are as follows:



confusion matrix_K-NN

Figure.3

TP (True positives) = 779

FP (False Positives) = 184

FN (False Negatives) = 211

TN (True Negatives) = 727

Classifier accuracy=(TP+TN)/
(TP+TN+FP+FN)=(779+727)/1901=79.2%

Applying Decision Tree classification:

Next, we applied Decision Tree classifier. The confusion matrix and accuracy of K-NN are given below:

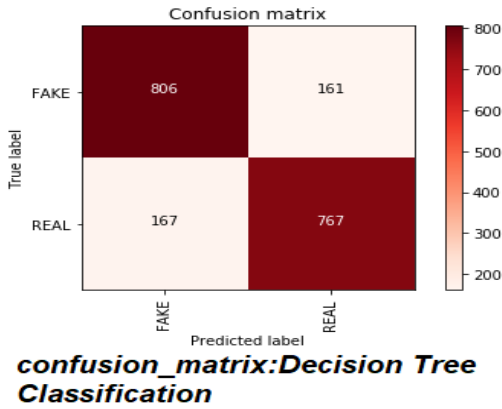


Figure.4

TP (True positives) =806

FP (False Positives)=161

FN (False Negatives)=167

TN (True Negatives) =767

Classifier accuracy= (TP+TN)/ (TP+TN+FP+FN) =
(806+767)/1901=82.7%

Applying Random Forest classification:

Next, we applied random forest classification algorithm.
Confusion matrix and accuracy are given below:

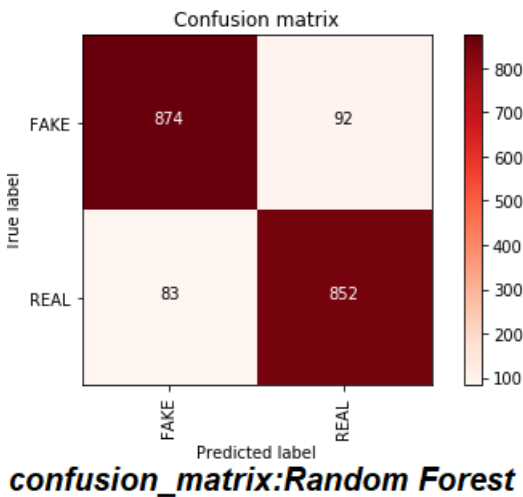


Figure.5

TP (True positives) =829

FP (False Positives) =110

FN (False Negatives) =183

TN (True Negatives) =779

Classifier accuracy= (TP+TN)/ (TP+TN+FP+FN) =
(874+852)/1901=90.7%

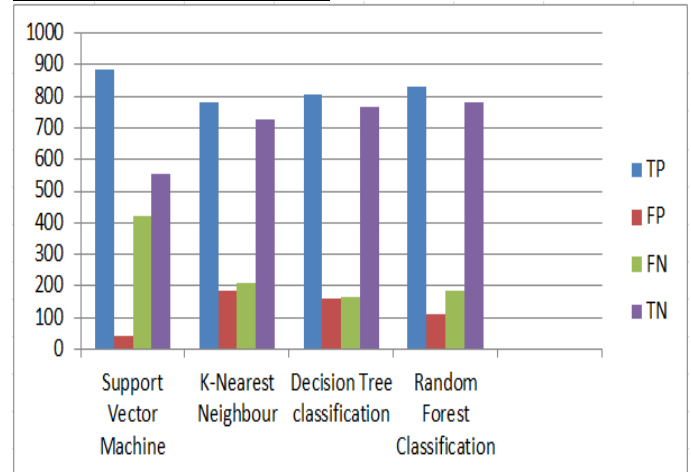
V. RESULTS

After implementing all classification algorithms, a comparison table is prepared for comparing Machine learning algorithms. Out of all classification algorithms, Random forest gives better accuracy.

Table1

Machine Learning Technique	TP	FP	FN	TN	Classifier accuracy
Support Vector Machine	884	43	421	553	75.5%
K-Nearest Neighbour	779	184	211	727	79.2%
Decision Tree classification	806	161	167	767	82.7%
Random Forest Classification	829	110	183	779	90.7%

Performance of algorithms:



performance of classification algorithms

Figure.6

Classifier Accuracy:

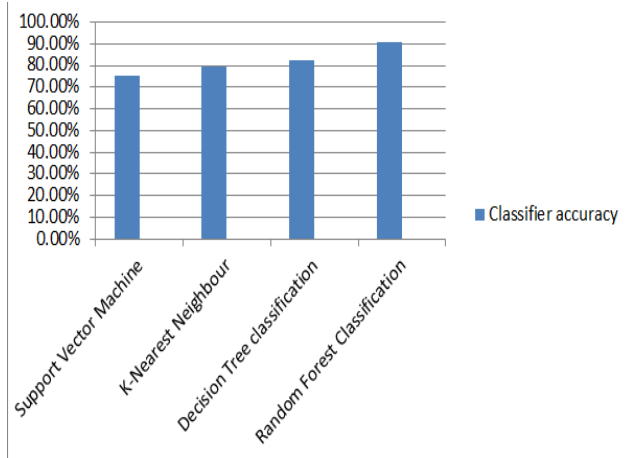


Figure.7

VI. CONCLUSION

In this paper, we build a model for fake news detection using machine learning algorithms. We evaluated four machine learning algorithms. For text processing, we applied machine learning NLP techniques. After that, we applied Support Vector Machine classification, K-NN classification, Decision Tree Classification, Random Forest Classification. SVM gives least accuracy. Random Forest gives best accuracy among all other classifiers.

REFERENCES

1. N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.
2. S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, Association for Computational Linguistics, 2012, pp. 171–175.
3. Junaed Younus Khan, Md. Tawkat Islam Khondaker, Anindya Iqbal and Sadia Afroz "A Benchmark Study on Machine Learning Methods for Fake News Detection", arXiv:1905.04749v1, [cs.CL] 12 May 2019
4. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community (ASIST 2015)*. Article 83, p. 4, American Society for Information Science, Silver Springs (2015)
5. Shlok Gilda, "Evaluating machine learning algorithms for fake news detection", 2017 IEEE 15th Student conference on Research and Development, INSPEC Accession Number: 17613664.
6. Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, :52(1), 1-4.
7. Hadeer Ahmed, Issa Traore I, and Sherif Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Technique", Springer International Publishing AG 2017. Traore et al. (Eds.): ISDDC 2017, LNCS 10618, pp. 127–138, 2017.
8. Schow, A.: The 4 Types of Fake News'. *Observer* (2017) <http://observer.com/2017/01/fake-news-russia-hacking-clinton-loss/>
9. Shlok Gilda, Department of Computer Engineering, Evaluating Machine Learning Algorithms for Fake News Detection, 2017 IEEE 15th Student Conference on Research and Development (SCORED)
10. Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," *SMU Data Science Review*: Vol. 1: No. 3, Article 10, 2018

AUTHORS PROFILE



Mr. A. Lakshmanarao, Assistant Professor from Raghu Engineering College, completed Bachelor of Engineering in CS&IT and M.Tech in Software Engineering. Currently pursuing Ph.D. in Machine Learning in Andhra University, Visakapatnam. His areas of interest are Machine Learning, Cyber Security and Deep Learning.



Mrs. Y. Swathi, Assistant Professor from BABA Institute of Technology & Sciences, Visakhapatnam, completed Bachelor of Engineering in CS&IT and M.Tech in Computer Science & Engineering. Her areas of interest are machine learning, Deep Learning.



Dr. T. Srinivasa Ravi Kiran, Assistant Professor & HOD, Department of Computer Science, P.B. Siddhartha College of Arts & Science Vijayawada completed Ph.D. in Acharya Nagarjuna University. His areas of interest are machine learning, Deep learning, Cyber Security