Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)
IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2

Performance Comparison of Machine Learning Classifiers for Fake News Detection

Smitha. N
Assistant Professor
Department of Computer Science & Engineering
CMR Institute of Technology
Bengaluru, India
smithan.ckm@gmail.com

Bharath .R
Department of Computer Science & Engineering
CMR Institute of Technology
Bengaluru, India
bhrr16cs@cmrit.ac.in

***Abstract*— Information sharing on the web particularly via web-based networking media is increasing. Ability to identify, evaluate and address such information is significantly important. Fake information deliberately created is purposefully or unintentionally engendered over the internet. This is affecting a larger group of society who are blinded by technology. This paper illustrates model and methodology to detect fake news from news article with the assistance of Machine learning and Natural language processing. In this proposed work different feature engineering methods like count vector, TF-IDF and word embedding are used to generate feature vector. Seven different Machine learning Classification algorithms are trained to classify news as fake or real and are compared considering accuracy, F1 Score, recall, precision and best one is selected to build a model to classify news as fake or real.**

Keywords— Fake news, SVM, Logistic Regression, XG-Boost, Gradient Boosting, Feature Extraction, NLP, Machine learning.

I. INTRODUCTION

The Internet is an abundance of data and exceptionally worthwhile for different reasons. Due to the overwhelming information available on the internet, one must be cautious about the originality [1]. To connect with family, friends, fellow workers the main medium is social media. Every user is sharing their feelings or information in the different forms like audio, video or text [2].

Fake information's are deliberately created and are purposefully or unexpectedly engendered over the internet. Creation and consumption of information over the internet have increased over time even if it's fake or real. Thus impacting groups of society who are large consumers of the internet and blinded by technology [3]–[6].

According to a survey, 77% of the USA population prefer to get news online over print media [7], [8]. Because of which it's highly important to conserve these data. In the current paper vulnerabilities of individuals and society because of fake news and also the extent it spreads is discussed. And also the requirement of mechanism to identify fake news and safeguard the community is discussed.

Since fake news tends to spread faster than the real news there is a need to classify news. In the proposed system the dataset used is from Kaggle website where real news and fake news are in two separate datasets and combined to one dataset and trained with different machine learning classification algorithms.

The fundamental target of this work is to propose a model which detects fake news by using three types of NLP text vectorization and applying on different ML classifiers and build a model which provides good results for classifying news as fake or real.

In this proposed system wellspring of the news are overlooked regardless of whether it was accounted from the web or in print and rather concentrate just on the substance matter.

II. BACKGROUND WORK

Social networking sites have changed the manner by which information is exchanged. Tacchini et.al has used a dataset consisting of 15,500 Facebook post. These were classified into two types, one based on logistic regression and another crowdsourcing algorithm and obtained 99% accuracy [9].

Rubin et al. by using SVM based model tested 360 news articles and obtained 90% precision, 84% recall and F1Score 87% [10].

A new classifier bilateral-weighted fuzzy support vector machine proposed and discussed how it exhibits its effectiveness and helpfulness [11].

Content has to be exceptionally ready before it could be utilized for predicting model. The content must be parsed to evaluate words and words should be ciphered as integer or floating-point values before giving it as input to a machine learning algorithm. Types of vectorization like a bag of words and word embedding are also be discussed [12].

Osisanwo et al. discussed various supervised machine learning classification methods and also compared and explained different attributes to evaluate the performance of the different method. Research is done using a dataset from the National institute of diabetes, digestive and kidney diseases [13].

Della Vedova et al. have proposed an ML model by considering news substance and social substance features and obtained higher accuracy and also deployed this method in a real-time application by implementing approaches with Facebook Messenger, Chatbot and obtained 81.7% accuracy [14].

Similarly, Granik et al. achieved 74% accuracy on test set data from Facebook news posts and they have mentioned that with AI approach accuracy could be increased [15].

Sen et al. have reviewed many supervised learning algorithms where they have discussed decision tree advantages and disadvantages. It is a straightforward strategy, quick, requires less information pre-processing, and can manage both categorical and numerical information but in some cases, this algorithm may prompt an intricate tree structure and it is a very unsteady model [16].

In another study, logistic regression and neural network was implemented for diabetes data obtained from HUSM and proved NN as the best predictive model. To achieve high accuracy, SVM has been suggested for future studies [17], [18]. In accordance with this, Burges explains the support vector training practical implementation. They have reviewed several arguments which support the high accuracy of SVMs [19].

Thus, with this background information in the proposed system, seven Machine learning algorithm is trained to check the originality of news.

III. METHODOLOGY

Different stages involved in the proposed system is as shown in the below Fig .1

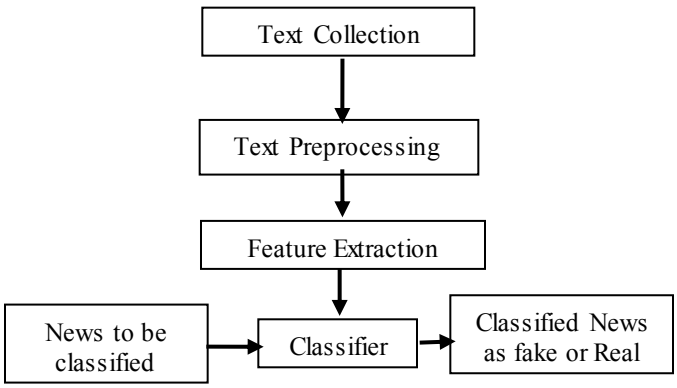

Fig. 1. System Architecture

A. Text Collection

The text collection process is carried out referring datasets collected from Kaggle [20]. Data are extracted from 244 websites. It consists of around 13000 posts recorded for 30 days. Where the training data set is 18574 post and Testing is 9149.

B. Text Pre-Processing

After the acquisition of content, pre-processing is carried out. Which includes the following steps:

- All letters in the document are converted to lowercase
- Numbers are removed
- Punctuations, accent marks are removed
- White spaces are removed
- Stop words are expelled

C. Feature Extraction

The text needs to be parsed to evaluate words and words should be encoded as integers or floating-point values before giving it into the machine learning algorithm [12]. Two types of vectorization methods used in the proposed system which include a bag of words and word embedding [21].

Bag of Words

Machine learning from text uses an approach called a bag of words which takes any text and counts the frequency of words after removing all the stop words [22]. After tokenization its need to be quantized and used on appropriate ML classifiers.

Countvectorizer: -

- The Count Vectorizer gives a basic method to both tokenize an assortment of content archives and fabricate a jargon of known words, yet additionally to encode new reports utilizing that jargon.
- An encoded vector is a comeback with a length of the whole jargon and a whole number mean the occasions each word showed up in the record.

TF-IDF:-

This is utilized to change over content to vectors thinking about the semantics of the word.Eq.3 is calculated by multiplying term frequency (1) and inverse document frequency (2).

TF = (Number of times term t appears in a document)/ (Number of terms in the document) (1)

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (1)$$

$$IDF = log(N/n) \quad (2)$$

where N is the number of archives and n is the number of archives a term t has appeared in word.

$$TF - IDF\ value\ of\ a\ term = (TF * IDF) \quad (3)$$

Word Embedding using Spacy: -

This follows a feature representation concept to produce a vector. Here low dimension, a dense matrix is achieved. Spacy a Natural Language Processing library used for word embedding to generate numeric vectors which represent a word. It is memory intensive and has undesirable effects.

D. Classifiers

In this proposed system different Machine learning algorithms are used to do the prediction. TF-IDF is used to generate vectors and all the listed algorithms applied to investigate the best calculation for counterfeit news recognition and the same procedure applied on word embedding and count vectorization.

1. Support Vector Machine

Machine to precisely classify SVM is used. Drawing decision boundaries are known as hyperplane which separates two classes. Unoptimized decision boundaries could result in misclassification, to overcome this SVM are considered as important by looking at extreme cases. Nonlinear SVM could be converted into linear by using some functions. Calculation creates the most ideal hyperplane which characterizes new data normally.

2. Logistic Regression

It is used for binary classification. Linear regression is used every time to create the best bit line for binary classification [23]. Logistic regression applied to problems where two classes would be linearly separable.

$$y = b0 + b1 * x \quad (4)$$

$$P = \frac{1}{1+e^{-y}} \quad (5)$$

$$\ln\left(\frac{P}{1-P}\right) = b0 + b1 * x \quad (6)$$

In Eq.6 b0 is the slope, x is a data point and b1 is intercepted. Equation (5) is the sigmoid function P, usually used to remove the effect of outliers [24].

3. Decision Trees

It's a predictive analysis. Here the graphical representation of all the possible solutions to a decision is made. A decision is based on some condition. Focal points are to incorporate word highlights, a direct partition of classes isn't required, effective treatment of exceptions, and a simple translation of the decision tree are required. Be that as it may, a decision tree would overfit when there are countless inadequate highlights, and consequently perform ineffectively on the testing information.

4. Random Forest

It's a troupe tree-based learning calculation. Builds multiple decision trees and merges them to produce more accurate and stable predictions. Trained with the bagging method. High variance obtained in the decision tree converted into low variance by using row sampling and feature sampling. Using hyperparameter the number of decision trees could be decided. It's an ensemble algorithm, which consolidates more than one calculation of the equivalent or distinctive kind for characterizing objects.

5. Gradient Boosting

Machine learning gradient boosting used for regression and classification. It's a boosting technique. Leaf represents an initial prediction, which is log(odds) which is used for classification, this is converted into a probability with logistic function (7).

$$Probability = \frac{e^{\log(odd)}}{1+e^{\log(odd)}} \quad (7)$$

6. XG-Boost

It's an extreme Gradient boosting. Designed to be used with large and complicated datasets. It's an ensemble method, regularized boosting by preventing overfitting. It's scalable in all scenarios [25]. It can handle sparse data and also parallel and distributed computation which makes learning faster and quicker [25].

Algorithm
Input: News Content
1.Convert text to lowercase.
2.Remove punctuations, digits, stop words from text.
3.Repeat
Input: Receive each news article
Calculate count vector for it.
Append the count vector to count_feature vector
Until the end of news articles.
4.Repeat
Input: Receive each news article
Calculate TF-IDF vector for it.
Append the TF-IDF vector to tfidf_feature vector
Until end of news articles.
5.Repeat
Input: Receive each news article
Calculate Spacy vector for it.
Append the Spacy vector to Spacy_feature vector
Until the end of news articles.
6.Parse count_feature vector, TFIDF_feature vector, Spacy_feature vector into classifier
Return feature vector gives us highest accuracy
7.Build model with the feature vector.
Output: Predicted label of news - Fake or Real

IV. RESULTS AND ANALYSIS

Different metrics used to evaluate ML classifiers. Accuracy compares predicted and actual labels. Precision used for information retrieval. Precision is calculated as out of the total positive results that are predicted by the model. Percentage of actual positive result which is relevant. Recall is True positive rate or True negative rate. F1score is a combination of precision &Recall

Table I. shows performance evaluation of ML classifiers with a count vector where different metrics are used. Table II. shows performance evaluation of ML classifiers with TF-IDF where different metrics are used, same with Table III. for word embedding.

Fig.2. shows a Neural network performs better with accuracy 0.94 and it is outperforming compared to other classifiers using count vectorizer. Fig.3. shows a comparison between different classifiers with TF-IDF vectorizer and SVM Linear is performing better than others with 0.94 accuracy.Fig.4. shows a comparison between different classifiers with word embedding, where the neural network performed better with 0.90 accuracy.

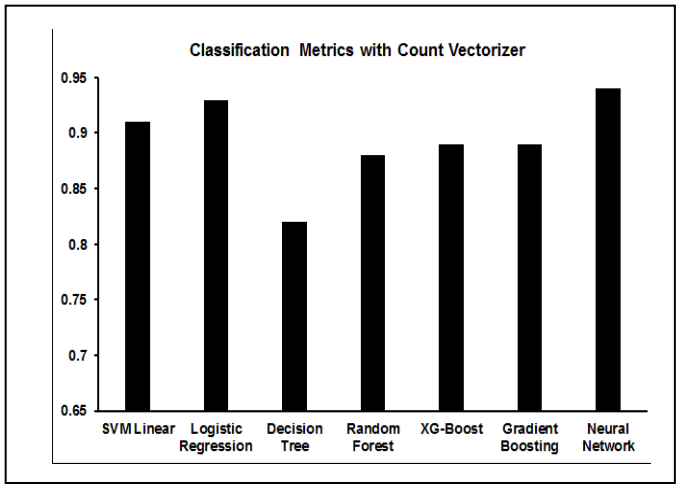

Fig. 2. Classification Metrics with Count Vectorizer

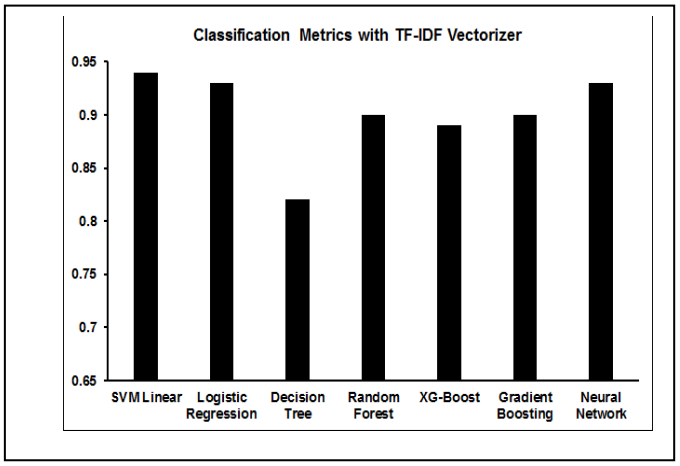

Fig. 3. Classification Metrics with TF-IDF Vectorizer

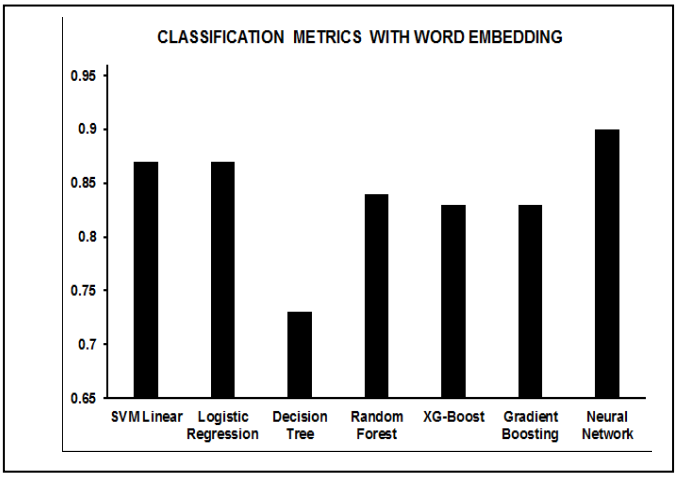

Fig. 4. Classification Metrics with Word Embedding

TABLE I. CLASSIFICATION METRICS WITH COUNT VECTORIZER

Classifiers	Performance evaluation			
	Accuracy	*Precision*	*Recall*	*Fscore*
SVM Linear	0.91	0.90	0.91	0.90
Logistic Regression	0.93	0.91	0.92	0.92
Decision Tree	0.82	0.80	0.81	0.80
Random Forest	0.88	0.94	0.79	0.86
XG-Boost	0.89	0.88	0.88	0.88
Gradient Boosting	0.89	0.89	0.88	0.88
Neural Network	0.94	0.94	0.93	0.93

TABLE II. CLASSIFICATION METRICS WITH TF-IDF VECTORIZER

Classifiers	Performance evaluation			
	Accuracy	*Precision*	*Recall*	*Fscore*
SVM Linear	0.94	0.93	0.93	0.93
Logistic Regression	0.93	0.93	0.91	0.92
Decision Tree	0.82	0.79	0.80	0.80
Random Forest	0.90	0.94	0.83	0.88
XG-Boost	0.89	0.89	0.88	0.88
Gradient Boosting	0.90	0.89	0.88	0.88
Neural Network	0.93	0.93	0.91	0.92

TABLE III. CLASSIFICATION METRICS WITH WORD EMBEDDING

Classifiers	Performance evaluation			
	Accuracy	*Precision*	*Recall*	*Fscore*
SVM Linear	0.87	0.88	0.83	0.85
Logistic Regression	0.87	0.88	0.82	0.85
Decision Tree	0.73	0.64	0.69	0.69
Random Forest	0.84	0.85	0.79	0.82
XG-Boost	0.83	0.84	0.76	0.80
Gradient Boosting	0.83	0.84	0.76	0.80
Neural Network	0.90	0.92	0.86	0.89

V. CONCLUSION

In this proposed system three different feature extraction methods like Count vectorizer, TF-IDF Vectorizer, Word Embedding has been used. And also, different classification algorithms are used.

By using classification algorithms, the highest accuracy obtained is with SVM Linear classification algorithm with TF-IDF feature extraction with 0.94 accuracy as shown in TABLE I, II, III. Even though the same accuracy with Neural Network with Count vectorizer is obtained, Neural Networks can take more time to train and it is complex. So, in this proposed system Linear SVM which is not so complex and takes less time to compute is considered.

VI. FUTURE WORK

In future, deep learning methods and sentiment analysis to categorize the news which may give high accuracy could be considered and further useful text like the publication of the news, URL domain etc., could be extracted for the process.

A dataset with a larger number of articles from different sources as it includes bigger jargon and more noteworthy substance could be used for more accuracy.

REFERENCES

[1] "The Advantages and Disadvantages of The Internet Essay," *Essays, UK. (November 2018)*. https://www.ukessays.com/essays/media/the-disadvantages-of-internet-media-essay.php?vref=1 (accessed May 07, 2020).

[2] Savita Kumari, "Impact of big data and social media on society," *Glob. J. Res. Anal.*, pp. 437–438, 2016.

[3] K. Nagi, "New Social Media and Impact of Fake News on Society," *ICSSM Proc.*, pp. 77–96, 2018, [Online].Available:https://ssrn.com/abstract=3258350.

[4] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science (80-.).*, 2018, doi:10.1126/science.aap9559.

[5] "http://www.businessinsider.in/Social-Big-Data-The-User-Data-Collected-By-Each-OfThe-Worlds-Largest Social-Networks-And-What-It-Means/articleshow/28969179.cms." .

[6] R. Palanisamy, N. Taskin, and J. Verville, "Impact of trust and technology on interprofessional collaboration in healthcare settings: An empirical analysis," *Int. J. e-Collaboration*, 2017, doi: 10.4018/IJeC.2017040102.

[7] "https://www.pewresearch.org/fact-tank/2019/09/11/key-findings-about-the-online-news-landscape-in-america/," 2019.

[8] "https://indianexpress.com/article/technology/social/ whatsapp-fight-against-fake-news-top-features-to-curb-spread-of-misinformation-5256782/."

[9] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it Hoax: Automated fake news detection in social networks," 2017.

[10] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News," 2016, doi: 10.18653/v1/w16-0802.

[11] S. Balasundaram and M. Tanveer, "On proximal bilateral-weighted fuzzy support vector machine classifiers," *Int. J. Adv. Intell. Paradig.*, 2012, doi: 10.1504/IJAIP.2012.052060.

[12] "https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/." .

[13] F. . Osisanwo, J. E. . Akinsola, O. Awodele, J. . Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.

[14] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," in *2018 22nd Conference of Open Innovations Association (FRUCT)*, May 2018, pp. 272–279, doi: 10.23919/FRUCT.2018.8468301.

[15] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, May 2017, pp. 900–903, doi: 10.1109/UKRCON.2017.8100379.

[16] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised Classification Algorithms in Machine Learning: A Survey and Review," in *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, J. Mandal and D. Bhattacharya, Eds. Singapore: Springer, 2020.

[17] S. Ahmad, T. N. T. Adli, A. A. Aziz, and J. Yacob, "Comparison in Neural Network Method to Find Comparing Logistic Regression and Neural Network Model for Type 2 Diabetes Mellitus among Obesity," *Int. J. Sci. Eng. Res.*, vol. 4, p. 2292, 2013.

[18] Y. Fei, J. Hu, W. Q. Li, W. Wang, and G. Q. Zong, "Artificial neural networks predict the incidence of portosplenomesenteric venous thrombosis in patients with acute pancreatitis," *J. Thromb. Haemost.*, 2017, doi: 10.1111/jth.13588.

[19] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, 1998, doi: 10.1023/A:1009715923555.

[20] "https://www.kaggle.com/mrisdal/fake-news." .

[21] P. . Joby, "Expedient Information Retrieval System for Web Pages Using the Natural Language Modeling," *J. Artif. Intell. Capsul. Networks*, vol. 2, no. 2, pp. 100–110, Jun. 2020, doi: 10.36548/jaicn.2020.2.003.

[22] Brownlee J., "How to Prepare Text Data for Machine Learning with scikit-learn," [Online]. Available: https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/.

[23] Xiaojin Zhu, "Text Categorization with Logistic Regression," *2007*, pp. 1–3, [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/cs838/LR.pdf.

[24] "https://www.javatpoint.com/logistic-regression-in-machine-learning." .

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016, doi: 10.1145/2939672.2939785.