Comparison of different machine learning models on various vectorization techniques to detect fake news

1st Vanita Jain
Bharati Vidyapeeth's College of Engineering
New Delhi
vanita.jain@bharatividyapeeth.edu

2nd Kanika Aapan
Bharati Vidyapeeth's College of Engineering
New Delhi
kanika1926@gmail.com

***Abstract*—As an increasing amount of our lives is spent interacting online over the internet, more and more people tend to seek and consume news from social media, news agency homepages,search engines,etc which can easily spread the fake news even faster than the real news. So there is a need for machine learning classifiers that can detect this fake news automatically. In this paper, we have applied various machine learning models using various feature vectorization techniques. The results reflects that the ensemble method works better than other machine learning models. Among all the ensemble methods ExtraTree classifier works best with accuracy 0.90.**

***Index Terms*—fake news detection , Logistic Regression, Decision Tree, Random Forest Classifier, Extra Tree Classifier, Gradient Boosting, KNN Classifier, Stacking Classifier, Count, TF-IDF Vectorizer.**

I. Introduction

Over the years we are progressing towards digitalization. In India, the number of internet users has been increased from 300 million in 2015 to 700 million in 2021. No doubt digitization increases our efficiency but we need to look at another side as well. We read a lot of information on the internet especially on social media feeds, which may appear true but often not[1]. This false information is deliberately created to misinform or deceive readers. Earlier we got the news from trusted resources like newspapers, media, journalists who first strictly verify the source and then publish the news for the readers. But due to the advancement of the internet and social media, the news gets shared among millions of people so rapidly without being verified. Usually, these fake news articles are created to either influence people's views, push a political agenda or cause confusion and can often be profitable for online publishers[2]. Therefore, researchers need to give attention to this problem. Detection of fake news is a big challenge as it is not an easy task. But first, let us define the term "Fake News". Fake news refers to news messages that contain incorrect or false information but do not report the incorrectness of information[3]. If the fake news is not detected in the early stages, it may spread to millions of people which may cause an irreversible effect on their life. There are some organizations that are dealing with this issue, verifying the news from the sources, but this problem is beyond the scope of human manual detection. So we need some efficient machine learning algorithms that can learn and perform by themselves.
In this paper, we experiment with various machine learning algorithms to detect fake news. Some of the traditional machine learning algorithms that we applied are Logistic Regression, KNN and ensemble methods like Random Forest and ExtraTree Classifier.

II. Dataset Description and Text Pre-processing

The dataset has attributes such as id, title, author, text, label.These attributes are described as follows:
`id`: Unique id for a news article.
`title`: The title of a news article.
`author`: Author of the news article.
`text`: The text of the article which could be incomplete.
`label`: that marks the article as potentially unreal.
Dataset is taken from Kaggle[4]
which consists of 20800 articles.
Our first step was to drop the rows from the dataset which

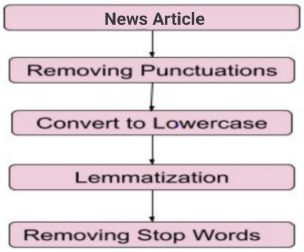

Fig. 1. Pre-processing steps

contains the null values in the text columns. The null values were also present in title and author column but as our experiment is based on the text column, we removed only that rows which contain null values in text columns. After removing the null rows,we got 20761 articles out of 20800.
To clean the text we applied various pre-processing techniques:
Removing punctuation: is done by removing the punctuation

characters from the sentences and filtering them out of the given text. After removing the punctuation marks we retained only the alphanumeric characters in the sentence[5].

Convert to lowercase: All the articles were read and converted to lower case simply by `string.lower()` function.

Lemmatization: converts the tokens to its base form or a valid dictionary word, that's why this is generally better than stemming. It is done using pos-tag lemmatizer from NLTK library[5,6].

Removing Stop words: used `nltk.stopwords()` library for removing stop words[5].

III. DATA ANALYSIS

With the help of the data analysis we can manually detect whether the news is fake or not.

Figure 2 shows the length of the fake and real articles. From this figure it can be analyzed that the length of fake articles is much more than the real articles.

Figure 3 and 4 shows the word cloud of real articles and fake articles respectively. From the word cloud it can be seen what are the most common words in the corpus of fake news and real news. It can be seen that trump, united states are some of the common words in the real articles and hillary clinton, people, say are some of the common words in fake articles.

Figure 5 and 6 shows the most 20 common bi-grams in the real and fake news article respectively[5].

Figure 7 and 8 shows the most 20 common tri-grams in the real and fake news article respectively[5].

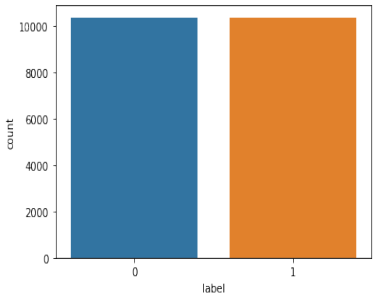

Fig. 2. Data is balanced

IV. METHODOLOGY

A. *Encoding the words*

We cannot directly feed the text data into the machine learning models. So, before training the data, we need to encode the text into number. We used four techniques to numerically represent the data:

*1) **TF-IDF Vectorizer**:* Tf-idf vectorizer converts a collection of raw documents into a matrix of tf-idf features[13]. The tf-idf score is the product the tf value and idf value for each term[5-8].

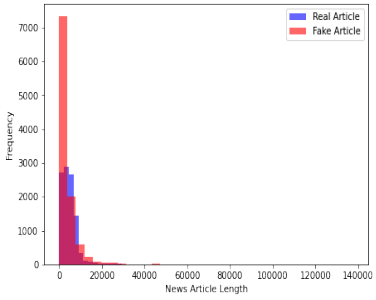

Fig. 3. Fake and Real text length

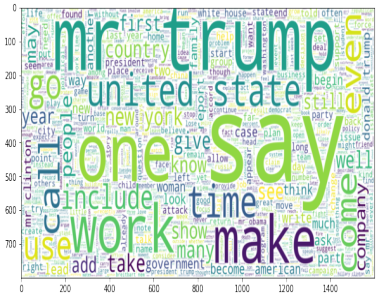

Fig. 4. Word Cloud for real articles

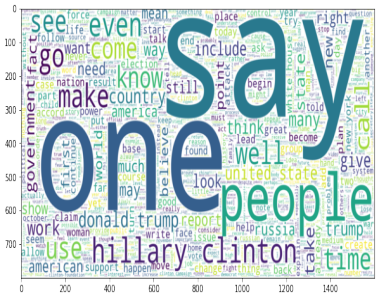

Fig. 5. Word Cloud for fake articles

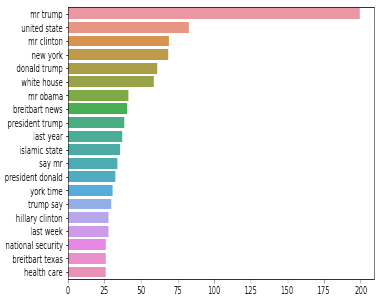

Fig. 6. Most common bi-grams words of real news article

- Term Frequency: Term frequency of term t is the number of times term `'t'` present in the document.
- Inverse Document Frequency: Document frequency is defined as the number of document `'d'` that contain the term `'t'`. Inverse document frequency is the reverse of document frequency.

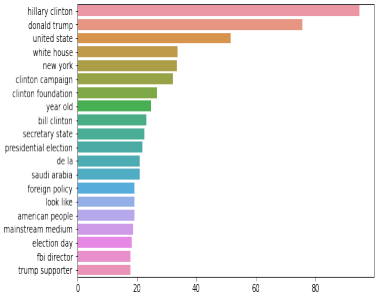

Fig. 7. Most common bi-grams words of fake news article

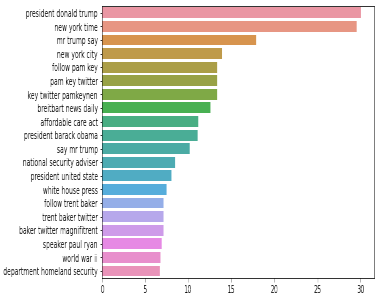

Fig. 8. Most common tri-grams words of real news article

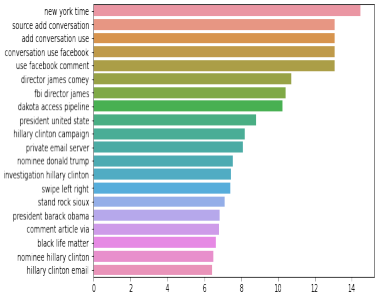

Fig. 9. Most common tri-grams words of fake news article

*2) **Count Vectorizer**:* Count Vectorizer converts the collection of text documents to a vector of token. Count Vectorizer uses the term frequency to represent a token in the documents[6-9].

*3) **Hashing Vectorizer**:* The Hashing Vectorizer applies a hashing function to term frequency counts in each document. This one is designed to be as memory efficient as possible. This vectorizer does not store token as the string rather this vectorizer applies the trick known as hashing trick to encode them as numerical indexes[7].

*4) **One Hot Encoding**:* One-hot encodes a text into a list of word indexes of size n where n is the size of the vocabulary. In One Hot Encoding we input a string of text and returns a list of encoded integers each corresponding to a word (or token) in the given input string[8].

B. Machine Learning models

*1) **Logistic Regression**:* Logistic Regression is a supervised algorithm that is used when the target variable is categorical. In Logistic regression the logistic function is used to model a binary dependent variable[10,13].

*2) **Multinomial Naive Bayes**:* Multinomial Naive Bayes classifier is used on very large dataset and at the same time it give good results also. It predicts the labels on the basis of conditional probability, and the Bayes theorem[11-13].

*3) **Decision Tree Classifier**:* The decision tree classifiers work on a series of questions for testing and conditions in a tree structure. It divides data based on that criteria at each level till it can classify a set of data, creating an explainable classification system[14,15].

*4) **Random Forest Classifier**:* Random forest is a supervised learning algorithm. The ”forest” it builds, is an ensemble of decision trees, mostly trained with the “bagging” method. It is the combination of learning models which increases the overall accuracy of the model[14,16].

*5) **Extra Trees Classifier**:* Extra Tree Classifier is an ensemble based learning technique. It fits various randomized decision trees on various random samples of the dataset to predict better accuracy and over-fitting also reduces[14,17].

*6) **Gradient Boosting Classifier**:* Gradient Boosting Classifier is one of ensemble boosting classifier. Gradient Boosting classifier is a strong classifier as it works by combining multiple poor classifiers so that the resultant accuracy obtained is high. Gradient Boosting give weights to the classifiers and train the data sample in each iteration to produce accurate predictions of unusual observations also[18].

*7) **KNN Classifier**:* K-Nearest Neighbour is one of the Machine Learning algorithms based on Supervised Learning technique. KNN algorithm stores all the available data and classifies a new data point based on the majority vote of its neighbors, with the case being assigned to the class most common among its K nearest neighbors measured by a distance function[19].

*8) **Stacking Classifier**:* Stacking Classifier is an ensemble based machine learning model. Two or more base models are used in the stacking classifier architecture. The model which is known as as a meta-model combine with the base model to make predictions. In stacking, the models are typically different and they fit on the same dataset. Using the above technique a single model is build by combining various machine learning models to predict good results. We stacked different classifiers such as Multinomial Naive Bayes, Extra-Tree classifier, Random forest and Logistic Regression[20]. Before using the stacking classifier ,tune the models on various parameters like depth and no. of estimators of the tree based classifier, tuning the number of iteration for each classifier, regularize parameter, smoothing parameters etc[21].

V. Results & Observations

F1 Score and accuracy measure is used as the evaluation metric to test which machine learning model performs best[22].

We observed that CountVectorizer gave better results than TF-IDF Vectorizer. It performed better,may be the count of

TABLE I
F1 SCORE

Classifier	Count Vector	Hash Vector	Tf-IDF Vector	One-Hot encoding
Logistic Regression	0.67	0.73	0.70	0.65
Decision Tree	0.66	0.65	0.67	0.70
Random Forest	0.86	0.87	0.85	0.70
Extra Tree Classifier	0.90	0.71	0.83	0.74
Gradient Boosting	0.67	0.66	0.67	0.71
KNN	0.76	0.74	0.71	0.65
Multinomial Naive Bias	0.57	0.55	0.58	0.66
StackingClassifier	0.66	0.65	0.63	0.69

TABLE II
ACCURACY SCORE

Classifier	Count Vector	Hash Vector	Tf-IDF Vector	One-Hot encoding
Logistic Regression	0.65	0.69	0.67	0.64
Decision Tree	0.64	0.63	0.64	0.67
Random Forest	0.85	0.86	0.84	0.67
Extra Tree Classifier	0.90	0.73	0.83	0.71
Gradient Boosting	0.65	0.64	0.65	0.68
KNN	0.70	0.64	0.58	0.65
Multinomial Naive Bias	0.58	0.58	0.59	0.64
StackingClassifier	0.64	0.63	0.61	0.67

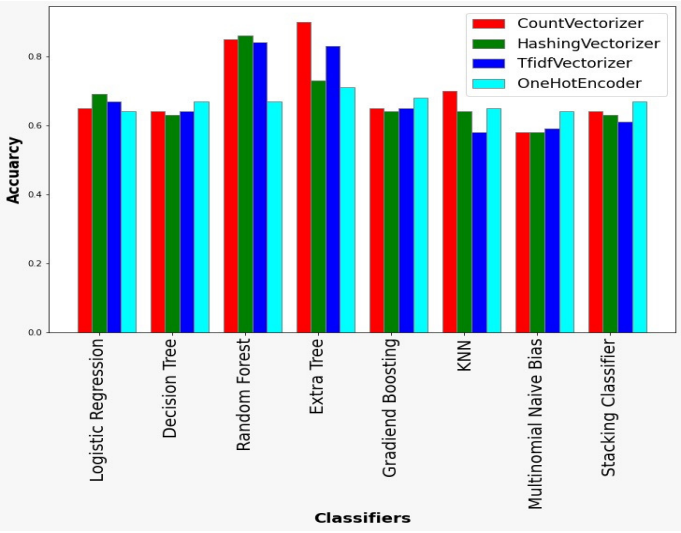

Fig. 10. Performance of all the classification techniques

words make the document more distinguishable between each other. From Table 1 we have seen that TF-IDF Vectorizer and CountVectorizer are performing similar. HashingVectorizer and One-Hot Encoding didn't perform well on the dataset.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have used different classifiers to identify the articles as fake or real. Firstly, we have applied various encoding techniques such as Tf-Idf Vectorizer, Count Vectorizer, Hashing Vectorizer and One-Hot encoding to convert the text data to numerical form. Then we have applied eight different machine learning models to detect the news as fake or real. We got the best results from **Extra-Tree Classifier** using the count vectorizer technique. In future we can try Deep learning techniques on the text and will try to get better results.

REFERENCES

[1] Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The spread of true and false news online." Science 359.6380 (2018): 1146-1151.

[2] Balmas, Meital. "When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism." Communication research 41.3 (2014): 430-454.

[3] Rubin, Victoria L., et al. "Fake news or truth? using satirical cues to detect potentially misleading news." Proceedings of the second workshop on computational approaches to deception detection. 2016.

[4] Dataset is collected from the given site : https://www.kaggle.com/c/fake-news/data/

[5] Ahmed, Hadeer, Issa Traore, and Sherif Saad. "Detecting opinion spams and fake news using text classification." Security and Privacy 1.1 (2018): e9.

[6] Oshikawa, Ray, Jing Qian, and William Yang Wang. "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770 (2018).

[7] Kaur, Sawinder, Parteek Kumar, and Ponnurangam Kumaraguru. "Automating fake news detection system using multi-level voting model." Soft Computing 24.12 (2020): 9049-9069.

[8] Huang, Jeffrey. "Detecting Fake News With Machine Learning." Journal of Physics: Conference Series. Vol. 1693. No. 1. IOP Publishing, 2020.

[9] Pal, Subhabaha, TK Senthil Kumar, and Sampa Pal. "Applying Machine Learning to Detect Fake News." Indian Journal of Computer Science 4.1 (2019): 7-12.

[10] Ahmad, Iftikhar, et al. "Fake news detection using machine learning ensemble methods." Complexity 2020 (2020).

[11] Aphiwongsophon, Supanya, and Prabhas Chongstitvatana. "Detecting fake news with machine learning method." 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE, 2018.

[12] Granik, Mykhailo, and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier." 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON). IEEE, 2017.

[13] Mahir, Ehesas Mia, Saima Akhter, and Mohammad Rezwanul Huq. "Detecting fake news using machine learning and deep learning algorithms." 2019 7th International Conference on Smart Computing Communications (ICSCC). IEEE, 2019.

[14] Hakak, Saqib, et al. "An ensemble machine learning approach through effective feature extraction to classify fake news." Future Generation Computer Systems 117 (2021): 47-58.

[15] Poddar, Karishnu, and K. S. Umadevi. "Comparison of various machine learning models for accurate detection of fake news." 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). Vol. 1. IEEE, 2019.

[16] Cusmaliuc, CIPRIAN-GABRIEL, LUCIA-GEORGIANA Coca, and A. D. R. I. A. N. Iftene. "Identifying Fake News on Twitter using Naive Bayes, SVM and Random Forest Distributed Algorithms." Proceedings of The 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR-2018) pp. 2018.

[17] Mahabub, Atik. "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers." SN Applied Sciences 2.4 (2020): 1-9.

[18] Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "Multiclass fake news detection using ensemble machine learning." 2019 IEEE 9th International Conference on Advanced Computing (IACC). IEEE, 2019.

[19] Kesarwani, Ankit, Sudakar Singh Chauhan, and Anil Ramachandran Nair. "Fake News Detection on Social Media using K-Nearest Neighbor Classifier." 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2020.

[20] Thorne, James, et al. "Fake news stance detection using stacked ensemble of classifiers." Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. 2017.

[21] Hellton, Kristoffer H., and Nils Lid Hjort. "Fridge: Focused fine-tuning of ridge regression for personalized predictions." Statistics in medicine 37.8 (2018): 1290-1303.

[22] Reis, Julio CS, et al. "Supervised learning for fake news detection." IEEE Intelligent Systems 34.2 (2019): 76-81.