ISSN No:-2456-2165

Classification and Analysis on Fake News in Social Media Using Machine Learning

Nihar Jain

Abstract:- Spread of fake news on social media has been a rising concern ever since the exponential growth of the internet. Propagation of fake news is rapid and can manipulate people'perception of reality effortlessly. Due to these deleterious effects of fake news, fake news detection has become essential and is gaining a lot of attention these days. Moreover, fake news is like real news in terms of structure and context. It becomes exceedingly difficult to detect fake news articles. Here, Machine Learning (ML) can play a significant role in classifying fake news. This study covers the classification of the fake news using machine learning models like Neural Network - Keras, Support Vector Machine (SVM), RandomForestClassifier and Logistic Regression. This study also covers the analysis of fake and real news articles that have been made to understand how these articles are structured. Next, there is a comparative analysis of the hate content in these articles with the publicly available datasets to understand the extent of hate content in fake news. Afterwards, there is the implementation of 4 ML models and the effect of performance for each model is measured with respect to the changes in the type of feature vector extraction techniques and the size of the dataset. Each of the ML models is evaluated in terms of its accuracy and there is further hyperparameter tuning performed to optimize the accuracy of the models. This study helps us in understanding the features of fake news articles and produces an optimal way to build a model to detect fake news that is propagating around us in social media.

I. INTRODUCTION

The internet is growing every day. It has become one of the best ways to share and collect information. With its rise over time, the use of social media is increasing exponentially, and people now are relying on it for getting information, facts, articles etc. People are now trusting information on social media and lack of in-depth knowledge on a subject gives online sources the benefit of doubt [1].

With the growing use of the internet, social media is becoming an integral part of our lives. It is used to share individual opinions, thoughts, facts. Information on social media is easily accessible and time saving. People share all kinds of information on social media. Social media for news consumption is a double-edged sword and it has become easy for people to spread malicious content.

The malicious content or 'fake news' is posted on social media for personal gains and circulates with a great speed over social media and results in a global impact. Some of the users might spread fake news unintentionally but most of the time there is a definite reason and plan behind its propagation. A small piece of information, when liked, commented, and shared, results in the widespread circulation of the information and can have ideological changes among people [21].

Fake news can cause certain biases in the minds of people and is dangerous for the people as well as the society. This news can manipulate people'perception of reality and has been used to influence politics, advertising industry and inculcate economic bias in the minds of the people. These fake news articles or tweets are structured in such a way that they seem legitimate to the minds of people. Some of the ill effects of social media include mob-lynching, demonstrations, protest, and riots [20].

Therefore, it becomes essential to identify this fake news on social media. Fake news identification takes time and resources. Manual identification of fake news by individuals takes a lot of time and is therefore not feasible. Intentional sharing of fake news is written by experienced users who structure the article in such a way that it becomes difficult to identify them as ethical or unethical. Machine learning and deep learning techniques are now applied to detect these types of articles and have turned out to be a good fit.

II. RELATED WORK

Fake news detection is becoming an emerging area for research. The current research areas for detection of fake news include text classification, sentiment analysis and implementation of blockchain based framework for reducing the spread of fake news. Fake news transmitted via social media has various aspects to it, for example, the number of followers of that user, the type of news or information tweeted / published, and analyzing their behavior and registration details. In [14] the authors have tried to find such aspects and have produced features such as the propagation of the article, the length of the tweet and the sentiment scores of the tweets.

Another research [15], based on structural properties of the social network, is used for defining a "diffusion network" which is the spread of a particular topic. This diffusion network together with other social network features can be helpful in the classification of rumors in social media with classifiers like SVM, random forest, or decision tree.

[7] proposed a model that combines 3 characteristics - the text of an article, the user response it receives, and the sources that are promoting it. This solution comprises 3 models - Capture, Score, and Integrate. The first model is based on response and text RNN is used to capture the user activity. The second model learns characteristics based on the behavior of the user. The first and the second model are combined with the third model to classify the fake news article.

In [8], authors have focused more on linguistic-based features such as total words, characters per word, frequencies of large words, frequencies of phrases (i.e., n-grams and bag-of-words approaches]), parts- of-speech (POS) tagging. Deep learning and machine learning techniques were also used to classify data by various authors. In [24], the authors used supervised machine learning algorithms for detecting fake news. For evaluation purposes, three different real-world datasets were used, and it was found that the Decision Tree Algorithm was better than the other algorithms in terms of accuracy, precision, and F-measure.

In [2], a hybrid model of LSTM and CNN was used to class fake news from Twitter posts. With the help of these advanced techniques and approaches, the model was explicitly able to identify patterns from the data and received an accuracy of 80%. Kaliyar et al. [10] went ahead with a pre-trained word embedding called Glove, which was later combined with a Convolutional Neural Network (CNN). The dataset used was a Fake News Dataset [11], and the proposed model performed exceedingly well.

III. DATASET

A. About the Dataset

The dataset taken is a collection of news articles that were found on various websites [18]. The articles consisted of 2045 news articles that were published by various authors on different sites. Our dataset consists of the following columns:

- *Author - the author of the news article*
- *Published - when were the articles published*
- *Title - title of the news article*
- *Text - content of the news article*
- *Language - language the article was written in*
- *Site_url - url of the site where the article was published*
- *Img_url - url of the image in the article*
- *Type - the type of the article (for example: hate, bias, conspiracy etc)*
- *Label - whether the article was real or fake*
- *Text_without_stopwords - text after the removal of stopwords*
- *Title_without_stopwords - text after the removal of stopwords*
- *Hasimage - whether the article contains only image or not (1 = presence of an image, 0 = absence of an image)*

B. Pre Processing

Before using the dataset for analysis and modeling, the data is preprocessed first. In the first step of this processing, all the words are put to lower case. Next, all the excess blank spaces in our articles are removed as these do not add any information to text processing. After this, all the punctuation is removed as they are not necessary for the model. Removal of stopwords is a crucial step in the preprocessing step.

Stopwords[19] are the common words that frequently occur but these do not carry any important meaning. Removal of such words helps us give more focus on the important words in the article that contribute to the article's meaning. Next, after the removal of the stopwords, lemmatization was performed. Lemmatization, in short, refers to the task of determining words that have the same root and converting them into their root form. Lemmatization takes into consideration the context of the words while converting them into their root form and therefore lemmatization has been used. After the basic preprocessing of the text, this dataset is then analyzed to discover patterns and trends within it and this analysis helps us understand the differences in fake and real news articles.

C. Data Analysis

The dataset consists of 1291 fake articles and 754 real articles. As mentioned above, the type column in the dataset represents the type of articles and sentiments present. There are a total of 8 categories in this type column. Below is the description of each type:

- *State: an article which talks about a particular state or nation*
- *Bs: an article which is deceptive in nature*
- *Bias: an article which has biased content*
- *Satire: an article which has exaggerated news*
- *Conspiracy: an article which consists of conspiracy*
- *Hate: an article which has hate present in its content*
- *Junksci: article which consists of outdated science news and is no longer relevant*
- *Fake: an article with fake sentiments*

Firstly, the type of articles for fake news as well as real news is looked upon. It has been observed that the type of the fake news articles is associated with conspiracy, fake, satire, junksci, bs whereas the real news articles were associated with bias, hate and state datatype. After further analysis, it has been found that the real news articles are most associated with bias and hate type whereas, fake news articles on the other hand, are most associated with the bs and conspiracy data type.

ISSN No:-2456-2165

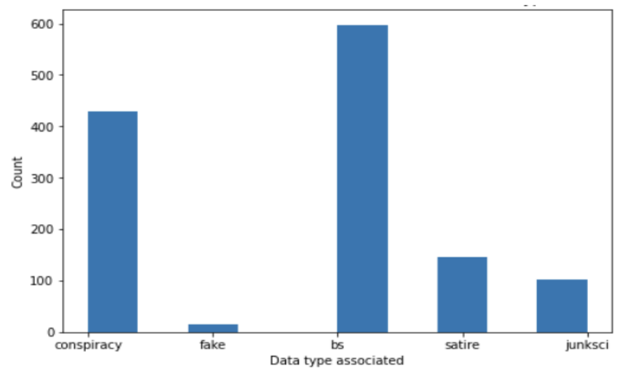

Fig 1 Number of Articles Associated with Fake News Data Types

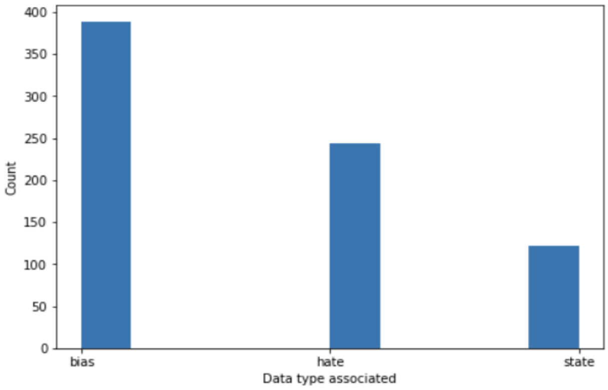

Fig 2 Number of Articles Associated with Real News Data Types

The results obtained suggest that real news articles are biased and have some sort of hate content in them. On the other hand, fake news articles are deceptive and have conspiracy in them. This makes sense as fake news is propagated to play with the minds of people and inculcate biases in them. The number of sites publishing fake news in our dataset was 59 while the number of sites publishing real news was 18. The total number of unique authors publishing fake news were 301 whereas it was 194 for that of real news. Further analysis is done to look at the type of articles these authors published. There is a specific focus on the most popular authors, and it was found that the results coincided with Fig 1 and 2.

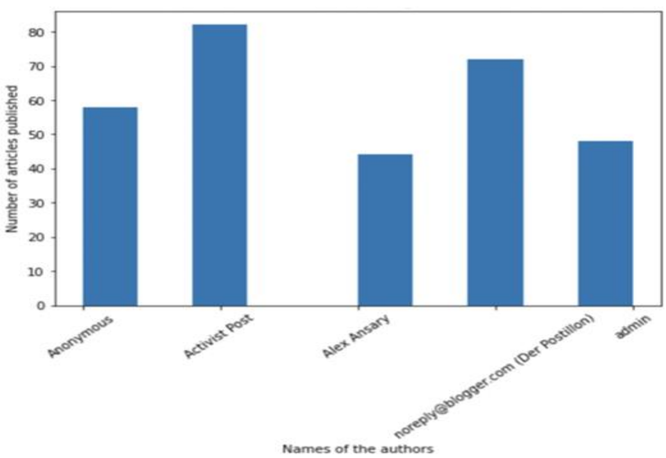

Fig 3 Top 5 Authors Publishing Fake News

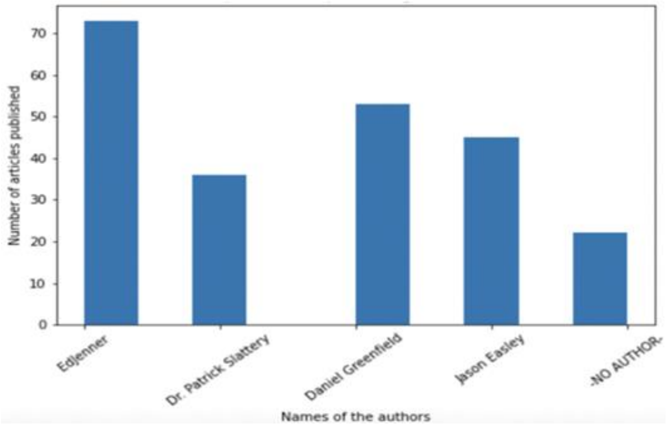

Fig 4 Top 5 Authors Publishing Real News

Fig 3 and Fig 4 show a pattern in the names of the author. It is observed that most of the authors who frequently published fake news articles used 'Fake name' and tried to hide their identity. From Fig 3, it is observed that 4 out of 5 authors had done this and this again correlates to the fact that there is intentional spreading of fake news. These authors deliberately spread fake news content to cause disorder in the society and therefore do not want to disclose who they are. However, such a thing is not observed in Fig 4. Another pattern found was that some of the authors who published only 1 or 2 fake articles did not actually hide their names and this may indicate that this spreading of fake news may be unintentional.

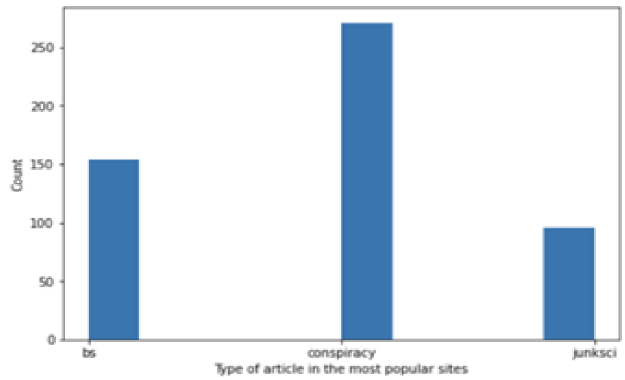

Fig 5 Type of Articles Found in Most Popular Sites Publishing Fake News

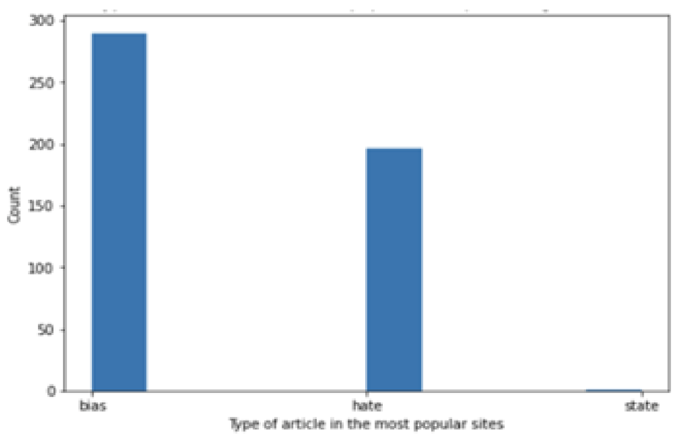

Fig 6 Type of Articles Found in Most Sites Publishing Real News

ISSN No:-2456-2165

Next, sites publishing fake news and real news are looked upon for analysis. For this, only the first 10 most popular sites are considered in both cases. For the websites publishing real news, related results are depicted in Fig 2, however, a contrast was seen for the websites publishing the fake news. There is more conspiracy content posted in the 10 most popular websites. This may suggest that though overall there is more bs content, websites which frequently publish fake news have more conspiracy articles in them to create biases in the minds of people.

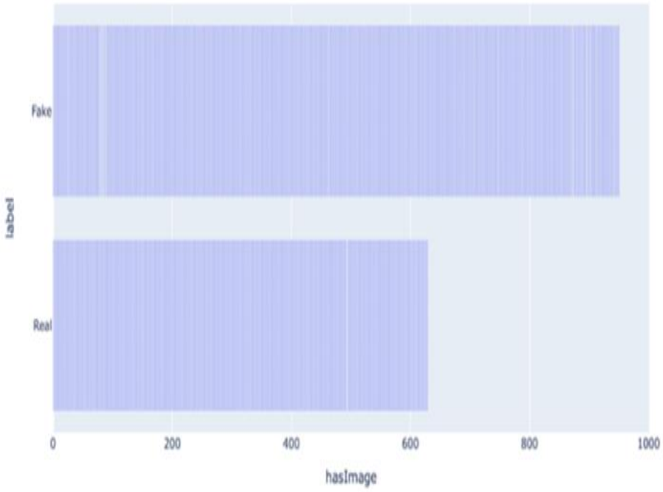

Fig 7 Articles Including Image VS Label

The dataset also consists of a column which indicates whether an image is present in the article or not. It has been observed that fake news articles have more images in them than the real ones. Fig 7 depicts the results obtained. There may be more images in the fake news articles to make them look more realistic and believable.

Fig 8 and Fig 9 show the text length of fake vs real articles and the average number of words per sentence in fake vs real articles, respectively. Usually, the fake news articles are shorter than the real ones. This observation is also supported by Fig 9 wherein the average number of words per sentence is less for fake news articles.

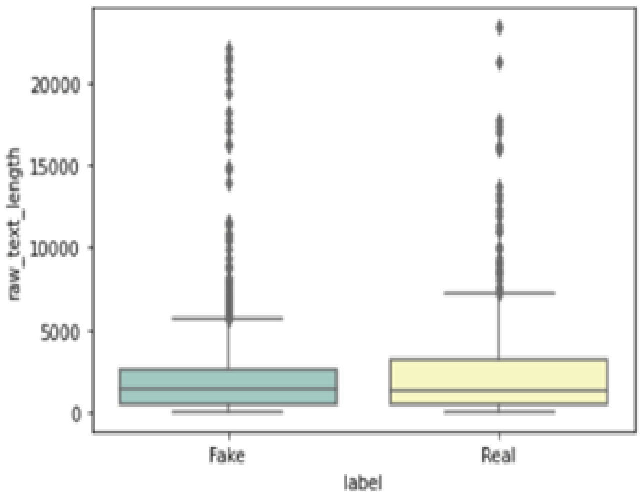

Fig 8 Raw Text Length of Fake VS Real Articles

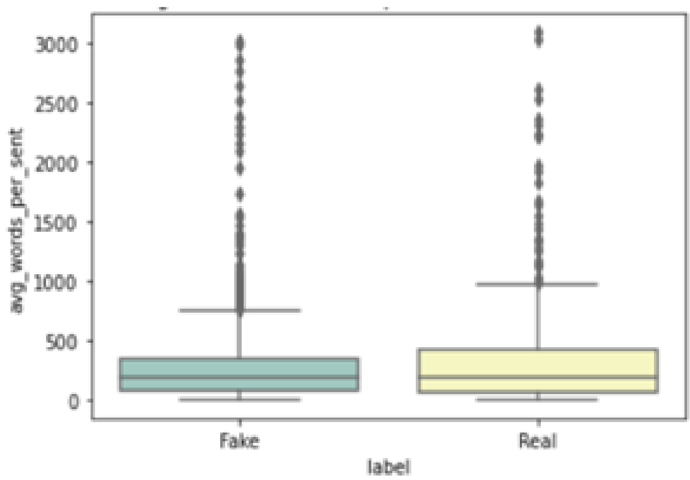

Fig 9 Average Number of Words Per Sentence in Fake VS Real Articles

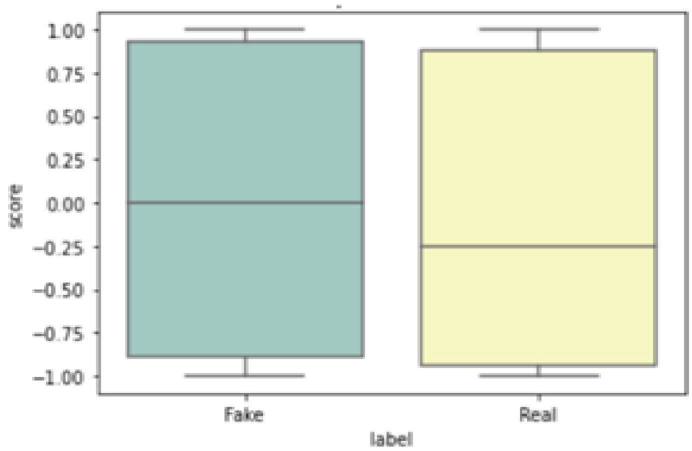

Fig 10 No of Articles Published Fake VS Real

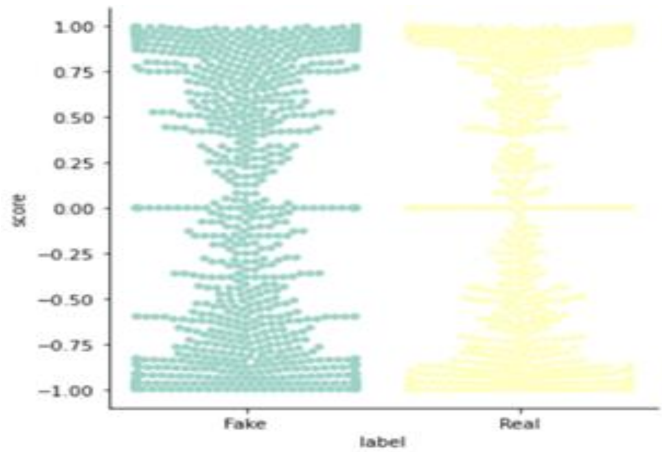

Fig 11 Senstiment Score Fake VS Real Articles

The sentiment scores for each article are calculated using theSentimentIntensityAnalyzer() function of nltk package. These sentiment scores indicate the polarity(positive/negative) of the article and the intensity of these sentiments. This function returns a dictionary of scores in 4 categories. These categories are negative, neutral, positive, and compound. For the analysis, the compound score of the sentiments is considered. The compound score is calculated by normalizing the scores obtained from the negative, neutral, and positive section. An article is classified with a positive sentiment if the compound score is greater than 0.3 and negative if the compound score is lesser than -0.3.

Real news articles are found to have more of a negative sentiment score and this correlates to our dataset wherein the types of real news articles are “hate and bias.” The cat plot (Fig 11) shows the distribution of sentiment scores for these articles. The average sentiment score of fake news articles is found to be 0.031 and that of real news is -0.081.

Features such as length of the articles, length of words in each article, number of positive words, number of negative words, average sentiment scores, most recurring positive and negative sentiments words in these articles are considered. However, most of these features did not show any striking difference between fake and real news articles. All these factors suggest that fake news articles are structured in a specific way and therefore classification of these articles becomes hard. In the next section, machine and deep learning models are used to classify news.

For further analysis of the hate content of fake and real news, a comparative analysis has been carried out. In this comparative analysis, 2 datasets are used [25][26] which contain hate and offensive speech. The idea here is to pick common words from the dataset to be analyzed and [25][26] dataset. Out of those common words, 20 popular positive and negative words are used from our dataset for analysis. A word is classified as positive if the compound score from the sentimental analysis is greater than 0.4 and it is negative if the compound score is calculated to be less than -0.4. Below are some of the popular words found in our dataset as well as in a hate speech dataset. The count of words below is derived from the dataset described in section 4.1.

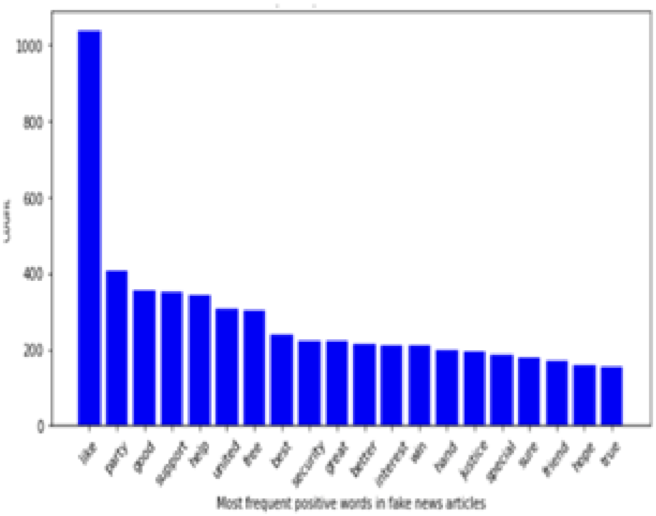

Fig 12 Count of Most Frequent positive Words In Fake News Articles

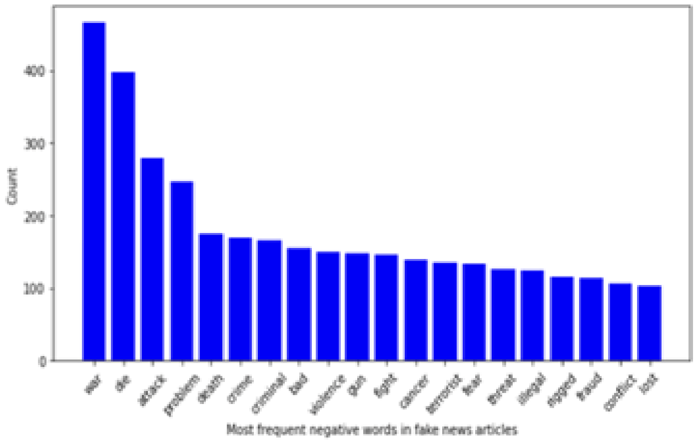

Fig 13 Count of Most Frequnet Negative Words in Fake News Articles

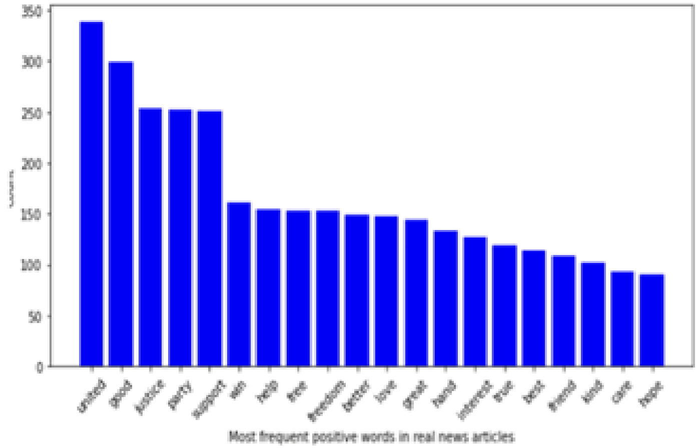

Fig 14 Count of Most Frequent Positive Words in Real News Articles

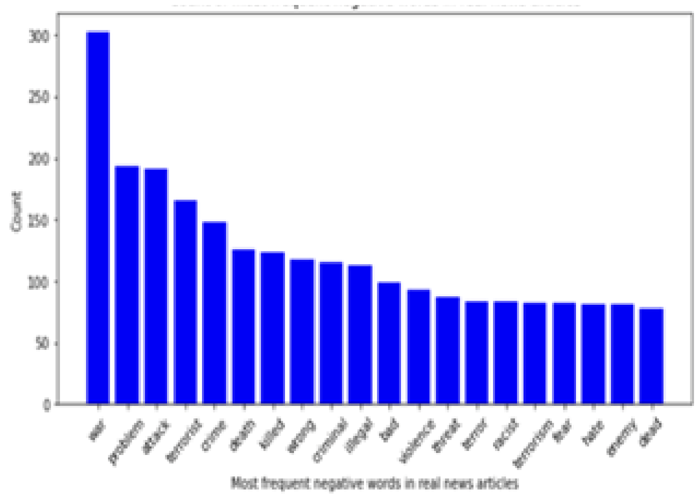

Fig 15 Count of Most Frequent Negative Words in Real News Articles

The above graphs depict that most of these words are repeated in fake news as well as real news articles. These negative words (Fig-13 and Fig-15) are found in hate speech datasets as well. It has been observed that the frequency of negative words in fake news articles is much more than the real news articles. This again correlates to the fact that fake news articles are propagated for the purpose of manipulating the minds of people and spreading hate in the society and creating bias. It is also observed from Fig-12 and Fig-14 that the frequency of positive words in fake news articles is higher than the real articles. The authors may include more positive words in fake news articles so that these articles seem legitimate.

This analysis shows that fake and real news articles are similar in structure, context, and sentiments. The authors of the fake news articles very subtly incorporate the fake information with frequent use of positive and negative words and propagate these articles / tweets through social media making them more legitimate.

IV. METHODOLOGY

A. Neural Network - Keras Model

Before any kind of modeling, the first step is to convert the textual data into its numeric forms. For this, several feature extraction methods are implemented to convert the data into numeric forms. In this study, there is an implementation of the Doc2Vec, TF-idf vectorizer, CountVectorizer techniques. The features extracted from each of the techniques are then applied to different machines and deep learning techniques. First the NN Keras model is implemented. Before passing any data, a feature vector is calculated for each of the articles under our dataset using Doc2vec model. While calculating the feature vectors for these articles, context of the article was taken into consideration. For each article, a numpy array of 100 elements is obtained. While calculating the feature vector for each article, only those words are taken into consideration that have occurred at least 5 times in our dataset just to avoid the effect of words that are not significant. The dimensions of each vector obtained are (100, 0).

After obtaining the feature vector for each article, the dataset is then split into training and testing data with the test size to be 0.2. The keras model requires the data type in tensor format and therefore both the training and testing data are converted into tensors. The neural network consists of 3 layers. The first layer is an input layer with 50 units. The activation function used is relu (rectified linear activation unit). The output from the first layer is passed to the second layer which consists of 20 nodes. Again, the activation function used in this layer was relu. The third layer, which was the output layer, consisted of 1 node and the activation function in the output layer was sigmoid.

B. Randomforestclassifier

Next, the RandomForestClassifier model is used for classifying fake news. The RandomForestClassifier is a classification algorithm which consists of several decision trees. This algorithm does not rely only on a single decision tree but takes into consideration the results obtained from multiple decision trees. The advantage of using multiple decision trees is that it leads to a higher accuracy and avoids the problem of overfitting.

The dataset requirements for this model are different from the previous model. The dataset as usual is split into training and testing sets but the size of the testing dataset was 0.1. The training datasets are then processed by the TfidfVectorizer converting the text into a matrix of TF-idf features. TF-idf features [23] is a statistical measure that considers the relevance of a word in the document. This extracts features from our dataset and assigns weights to the words in the dataset. The extracted features are then passed to our model. The 3 parameters used in the model are n_estimators, learning rate and the algorithm. The value of n_estimators is kept as 1, learning rate is chosen to be 0.9 and the algorithm chosen is SAMME. This model has achieved an accuracy of 0.771 with these parameters.

C. Support Vector Machines (Svm)

Next, the SVM model is used to classify fake news articles. The SVM model, in classification problems, creates a line which separates the data points into classes. It works by mapping a n-dimensional feature space in such a way that these data points can be classified. This algorithm determines the best line or decision boundary between vectors that can be used to divide the data points into categories.

TF-idf features, calculated in section 5.2, are then passed on to the model as the training data. The parameters used in modeling are - kernel, C (avoids misclassification) and degree. The kernel is set to be linear. The value of C is 90 and the degree is set to 2. The accuracy of this model is 0.7707.

D. Logistic Regression

Logistic regression is also implemented for text classification. Logistic regression is an algorithm that is used to predict a binary outcome (in this case Fake/ Real). The CountVectorizer is used for feature extraction of articles. CountVectorizer works on the principles of frequency. CountVectorizer converts a collection of text documents to an array of token counts. The feature vectors obtained are passed to the Logistic Regression model as the training dataset. The accuracy of the Logistic Regression is 0.7382.

In all four models discussed above, only 1 type of feature extraction method is used in the first stage. In the second stage, each model is given the feature vectors from all the techniques discussed in section 5.1 - 5.4. Below table 1 captures the accuracies of all four models.

ISSN No:-2456-2165

Table 1 Accuracy Obtained from the Different Models (Original Dataset)

	Model used for classification			
Feature Extraction Method	**NN - KERAS**	**SVM**	**Logistic Regression**	**RandomForestClassifier**
Doc2Vec	0.5978	0.6319	0.6205	0.5550
TF-idf	**0.7482**	**0.7707**	**0.7565**	0.7208
Countvectorizer	0.6502	0.7050	0.7382	**0.7320**

Based on the results obtained from Table 1, it is observed that the TF-idf vectorizer performed the best across all the models. The TF-idf vectorizer can capture more variance in the dataset and this results in higher accuracy of the models. The CountVectorizer performed moderately while the Doc2Vec model did not perform so well.

E. Upsampling The Dataset

Our dataset is small, consisting of only 2045 samples. The models implemented above have performed well because the size of the dataset used is very small. To create a better classification model, the dataset was upsampled. Upsampling of the dataset is done by adding more data from another dataset.

The dataset added is sourced from kaggle [22] consisting of 6335 samples. Out of these 6335 samples, 3164 are fake news articles and 3171 are real news articles. This dataset is processed in the same way as discussed in section 4.2.

After combining the 2 datasets, the resultant dataset consists of a total 8380 samples with 4455 fake articles and 3925 real news articles. Features are extracted using the techniques discussed in section 5.1, 5.2, 5.4 and the extracted features are again passed to the models as training dataset. Below are the accuracies of the models obtained with the upsampled dataset.

Table 2 Accuracy Obtained from the Different Models (Upsampled Dataset)

	Model Used - Upsampling			
Feature Extraction Method	**NN - KERAS**	**SVM**	**Logistic Regression**	**RandomForestClassifier**
Doc2Vec	**0.8645**	0.8015	0.8102	**0.8250**
Tf-idf	0.8562	**0.8512**	0.8447	0.7860
Countvectorizer	0.7906	0.8504	**0.8529**	0.7991

It has been observed that as the dataset size increases, the models start performing better. Neural Network, SVM and Logistic Regression have been able to achieve a higher accuracy and result in a better classification of the fake news articles.

V. EVALUATION AND HYPERPARAMETER TUNING

In the evaluation stage, receiver operating characteristic (ROC) curve and area under the ROC Curve (AUC) scores are used to evaluate the models. For each model, only 1 out of 3 feature extraction methods are considered for the evaluation stage. ROC curves are plotted from them. The ROC curve shows a trade-off between true positive rate (TPR) and false positive rate (FPR). The AUC is another measure to test the performance of a model; the higher score of AUC results in higher performance of the model.

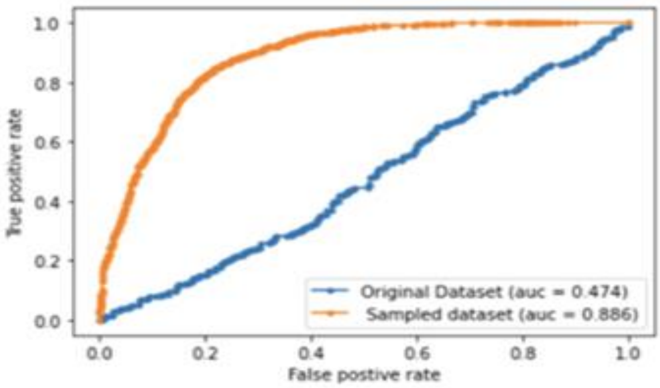

Fig 16 Roc Curve of NN Model With Doc2Vec

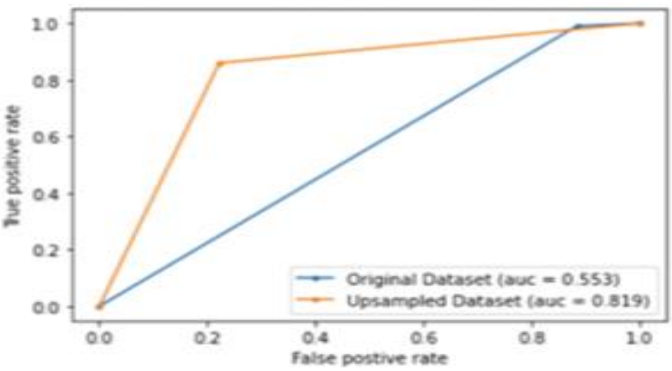

Fig 17 Roc Curve Random Forest Classifier With Count Vectorizer

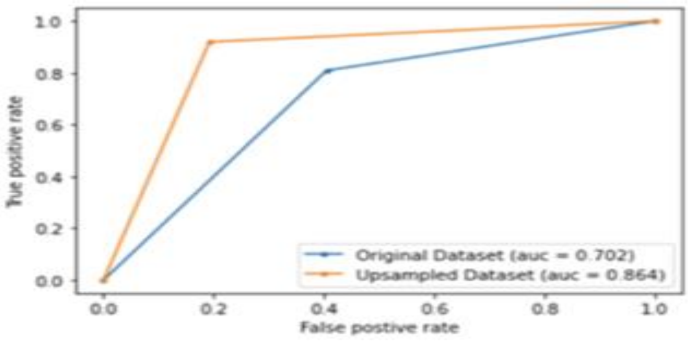

Fig 18 Roc Curve of SVM With TF-IDF vectorization

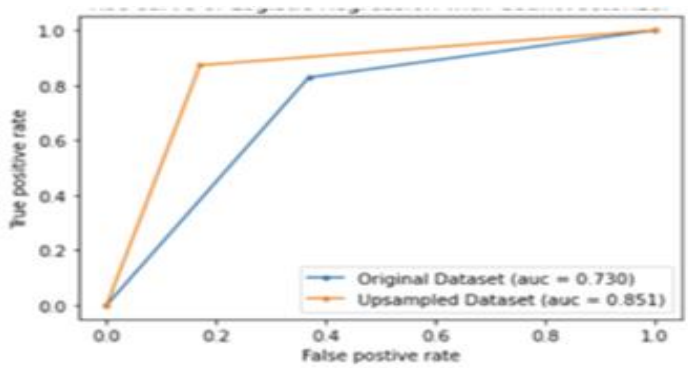

Fig 19 Roc Curve of Regression with Count Vectorizer

From the above ROC curves and AUC scores, it can be concluded that the models have significantly improved after the upsampling of the dataset. The neural network showed the best results wherein the model accuracy went up from 59.78% to 86.45%. RandomForestClassifier as well as Logistic Regression also showed significant improvement after the upsampling of the dataset. The SVM model also performed well in both the datasets indicating that it can be used for small-sized datasets.

A. Hyperparameter Tuning

To further optimize our models, hyperparameter tuning is performed for the models. The models chosen for hyperparameter tuning are RandomForestClassifier with CountVectorizer and SVM with TF-idf vectorization.

➢ RandomForestClassifier:

Hyperparameter tuning for the RandomForestClassifier is done with the help of the Grid search method. The Grid search method tries to compare different models and finds out the value of parameters that gives us the best results. For each parameter, different variations of parameters are used in the algorithm and after exhaustive search, the algorithm returns the value of parameters leading to the best accuracy of the model.

Grid search is applied to the RandomForestClassifier with three optimized parameters n_estimators, learning rate and the algorithm. Below are the results of the model after the hyperparameter tuning.

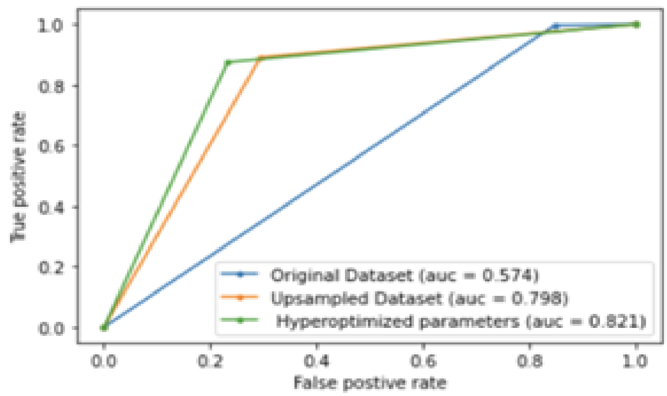

Fig 20 Roc Curve of Random Forest Classifire with Count vectorizer

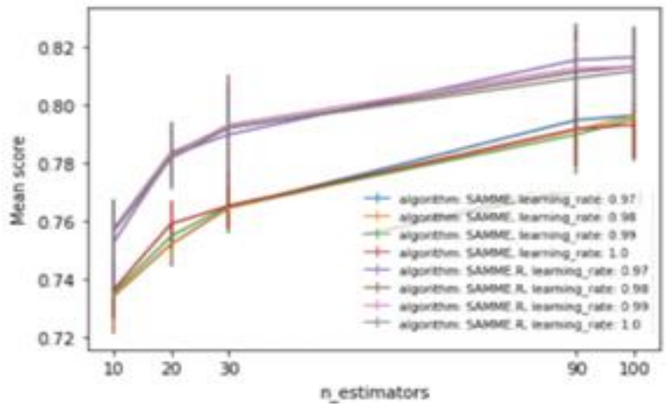

Fig 21 Gird Search Result

The Accuracy of the model after tuning was 0.8429. The higher the mean score, the better the model accuracy and precision. From Fig 21, it is observed that as the value of n_estimators is increased, model accuracy also increases. The model gave the best results when the value of n_estimators is 100, learning rate is 0.97 and the algorithm used is SAMME.R.

➢ *Support Vector Machines (SVM):*

Hyperparameter tuning for SVM is again done through the grid search method. There are 3 parameters to be optimized- C, kernel and degree. C was set as 16, kernel as rbf and degree as 10. The accuracy of the model after tuning was 0.8757.

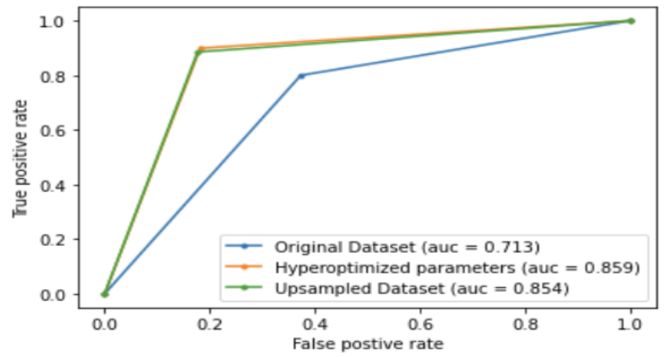

Fig 22 Roc Curve of SVM with Countvectorizer

The ROC curve of the optimized model was slightly more than the original one. There is approximately 1% increase in the accuracy of the model. Although, hyperparameter tuning was not so effective in case of SVM, from the above 2 examples, we can conclude that hyperparameter tuning helps in increasing the accuracy of the models resulting in better classification of fake news.

VI. CONCLUSION

Fake news articles are published to create an imbalance in the society. Fake news and real news articles have similar characteristics and therefore, it becomes difficult to classify the fake news manually. A significant amount of hate speech content is present in fake news articles, and this is done to create biases in the minds of people. Most of the fake news articles have images and more positive words to make the article more realistic. The propagation of these fake news articles is planned and can cause a global impact. Thus, the identification or propagation of these articles is essential, and this can be achieved with the help of machine learning models with greater accuracy.

In this study, four machine learning models were implemented with different feature extraction methods. All these feature extraction methods showed different results with different models and the models performed better by increasing the size of datasets. Hyperparameter tuning further helped us in increasing the model accuracy. All these models can be used for the identification and classification of fake news. Usually, these models perform better with larger datasets. However, some models like SVM and logistic regression perform well even with smaller datasets. Though, Neural Networks do not give promising results with smaller datasets and specifically require larger datasets. With the smaller dataset, it was seen that the TF-idf vectorizer was better able to extract the features in the dataset and most of the models worked best with this feature extraction method. However, the Doc2vec technique showed better results by increasing the size of the dataset.

For creating an optimal classification model, the first step is to choose the right model and the feature extraction techniques that are used for classification purposes. The second step includes the passing of feature vectors to the model as the training data. These feature vectors are extracted using the different feature extraction techniques. Applying datasets upsampling and hyperparameter tuning techniques give promising results in improving the accuracy of machine learning models. Implementation of these models can help in identifying fake news in real life scenarios.

REFRENCES

[1]. https://www.divaportal.org/smash/get/diva2:1479925/FULLTEXT01.pdf
[2]. https://dl.acm.org/doi/10.1145/3217804.3217917
[3]. J. Curci, K. Khara, A. Pillai, and R. Qin, “Fake News Detection,” Boston University, 2018.
[4]. Kaggle, “Fake News , ” 2018 . [Online]. Available: https://www.kaggle.com/c/fake-news/data. [Accessed: 07-Mar-2020]
[5]. Granik, Mykhailo ,Volodymyr Mesyura.(2017). “Fake News Detection using Naive Bayes Classifier.” *IEEE First Ukraine Conference onElectrical and Computer Engineering (UKRCON)*:900-903.
[6]. S. Gilda. (2017). “Evaluating Machine Learning Algorithms for Fake News Detection,” *IEEE 15th Student Conference on Research and Development (SCOReD)*, Putrajaya: 110-115.
[7]. RUCHANSKY,Natali,SEO,Sungyong,etLIU,Yan.Csi :Ahybriddeep model for fake news detection. In : Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017. p. 797-806.
[8]. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
[9]. H. Ahmed, I. Traore, S. Saad, Detection of Online Fake News using N-Gram Analysis and Machine Learning Techniques, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2017, pp. 127–138, http://dx.doi.org/10.1007/978- 3- 319- 69155- 8_9.
[10]. R.K. Kaliyar, A. Goswami, P. Narang, S. Sinha, FNDNet – A deep convolutional neural network for fake news detection, Cogn. Syst. Res. 61 (2020) 32–44, http://dx.doi.org/10.1016/j.cogsys.2019.12.005.
[11]. UTK Machine Learning Club, Fake news, 2018, https://www.kaggle. com/c/fake- news/overview (accessed June 3, 2020).
[12]. P. Bahad, P. Saxena, R. Kamal, Fake news detection using bi- directional LSTM-recurrent neural network, Procedia Comput. Sci. 165 (2019) 74–82, http://dx.doi.org/10.1016/j.procs.2020.01.072.
[13]. Jruvika, Fake news detection, 2017, https://www.kaggle.com/jruvika/ fake- news-detection/version/1 (accessed June 3, 2020).
[14]. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684 (2011)
[15]. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining, pp. 1103–1108. IEEE (2013)

[16]. Zhang, J., Dong, B., Philip, S.Y.: Fakedetector: effective fake news detection with deep diffusive neural network. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1826–1829. IEEE (2020)

[17]. Kaliyar, R.K., Goswami, A., Narang, P., Sinha, S.: FNDNet-a deep convolutional neural network for fake news detection. Cogn. Syst. Res. 61, 32–44 (2020)

[18]. Dataset: https://github.com/jnihar/classification

[19]. Stopwords:[line number 73] https://www.nltk.org/nltk_data/

[20]. Wani, A., Joshi, I., Khandve, S., Wagh, V., Joshi, R. (2021). Evaluating Deep Learning Approaches for Covid19 Fake News Detection. In: Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds) Combating Online Hostile Posts in Regional Languages during Emergency Situations. CONSTRAINT 2021. Communications in Computer and Information Science, vol 1402. Springer, Cham. https://doi.org/10.1007/978-3-030-73696-5_15]

[21]. Kumar, Srijan, and Neil Shah. "False information on web and social media: A survey." *arXiv preprint arXiv:1804.08559*(2018).

[22]. https://www.kaggle.com/datasets/jillanisofttech/fake-or-real-news

[23]. https://medium.com/acing-ai/what-is-tf-idf-in-feature-engineering-7f1ba81982bd

[24]. Ozbay FA, Alatas B (2020) Fake news detection within online social media using supervised artificial intelligence algorithms. https://doi.org/10.1016/j.physa.2019.123174. (https://www.sciencedirect.com/science/article/pii/S0378437119317546)

[25]. https://www.kaggle.com/datasets/kinguistics/election-day-tweets?select=election_day_tweets.csv

[26]. https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset