International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-09, Issue-12, Mar 2024

Comparative Analysis of NLP Techniques and Machine Learning Algorithms for Fake News Detection

[1]Mr. Satyam Maurya, [2]Mr. Ayush Gupta, [3]Dr. Kirti Wanjale

[1,2]Department of Information Technology, [3]Associate Professor, [3]Department of computer engineering, BRACT's Vishwakarma Institute of Information Technology, Pune, Maharashtra, India. [1]satyam.22011203@viit.ac.in, [2]ayush.22011191@viit.ac.in, [3]kirti.wanjale@viit.ac.in

Abstract – As we all know how rapidly social media is emerging and in today's world almost everyone has access to internet, because of this fake news spreads very quickly. Fake news has a big impact on our social lives, mostly in politics and education.

In this research study, we embarked on the task of converting textual news headlines into vectorized representations using Natural Language Processing (NLP). Two NLP techniques, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), were explored and compared to determine their respective efficacy in fake news detection.

We evaluated the performance of several machine learning classification algorithms, including Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine. Our aim was to identify the most effective method for identifying fake news.

Our study concludes that TF-IDF is more effective as compare to BoW, and random forest was giving maximum accuracy among all the machine learning models we used.

This study provides insight into the advantages and disadvantages of various NLP techniques and machine learning algorithms, which advances the field of fake news detection.

Keywords - Fake news detection, NLP, BoW, TF-IDF, SVM, Random Forest, Logistic Regression, and Naïve Bayes.

I. INTRODUCTION

Our lives now centre mostly around social media. That is where most fake news gets spread. A serious issue affecting politics, the financial system, democracy, businesses, and education is fake news. Social media, which occasionally spreads misleading information and leads others to believe it, has earned people's trust. It is getting more difficult to distinguish between fake and real news, which causes misunderstandings and confusion. It is difficult to identify fake news manually. It requires in-depth knowledge of the subject. But it is now simpler to produce and distribute fake news thanks to recent developments in computer science. Determining the accuracy of the information remains challenging, though. Businesses may be impacted when false information about their goods circulates [18].

In this study, our goal was to use a technique known as Natural Language Processing (NLP) [15] to convert text-based news headlines into numerical forms. Bag of Words and Term Frequency-Inverse Document Frequency are NLP techniques that we used. The goal was to determine which one was more capable at identifying false news. Additionally, we compared different machine learning methods. We used Naïve Bayes, Logistic Regression, Random Forest, and Support Vector Machine [16][17]. Our aim was simple: to figure out which one was the best at telling fake news apart from the real stuff.

So, to sum it up we first transformed news into numbers with NLP after that we used various machine learning algorithm to decide which one is giving better accuracy.

The format of this paper is as follows: Related work is covered in Section 2, the dataset used is explained in Section 3, our model's implementation is presented in Section 4, results are discussed in Section 5, Section 6 concludes the paper and in section 7 we discussed future scope of the research.

II. LITERATURE SURVEY

In [1] dataset used was combination of two datasets, FakeNewsNet dataset and McIntire Dataset, feature extraction was done using linguistic features (stylometric)

DOI : 10.35291/2454-9150.2024.0065

and word vector features (TF and TF-IDF) and finally various ML models were applied on it like NB, SVM, KNN, LR, RF. The results revealed that all the machine learning models performed effectively when stylometric features were employed.

In [2] data was extracted from a news website, and the study employed the Naive Bayes and SVM algorithms. The results were compared with pre-existing models for evaluation. But the problem with this model was that since the news was taken from news website itself, so there is high probability that the news was authentic.

In [3] TF and TF-IDF were used for NLP tasks, and the algorithms Naive Bayes, SVM, and Passive Aggressive Classifier were used. These results were compared with existing models, among the NB, SVM and PAC, SVM was yielding the highest accuracy.

In [4] dataset used was a public dataset published by Signal Media for research purpose. NLP techniques used included Bigram Term Frequency-Inverse Document Frequency and Probabilistic Context-Free Grammar (PCFG). The union of these techniques was also examined. Various models, including Random Forest and Support Vector Machine, were used. It was observed that the model trained using TF-IDF features performed the best, while the PCFG-based model did not significantly enhance predictive value. The findings indicated that SVM, in combination with TF-IDF feature selection, yielded the highest accuracy.

Authors of [5] have made comparison between ML and DL for this problem. for ML models, methodology was like [4] and for DL model the authors have implemented LSTM with various variations like LSTM, LSTM (with dropout regularization) and LSTM with CNN. The accuracies achieved with ML models was near about 70% and with LSTM they were able to achieve accuracy near about 80%.

Authors of [6] merged two datasets to get a new dataset namely, Getting Real about Fake News and all the news dataset. They used ngram, bow and TF-IDF for NLP, kept n = 2 for ngram also they studied effect of various features like sentiment, data, source, and author on accuracy. After considering the features to be used they applied SVM algorithm and changed their kernel type (linear, Radial basis, polynomial) to see which kernel was best performing.

For [7] dataset was taken from Kaggle, authors of [7] used word embedding (one hot encoding) for NLP and for model they used LSTM. Accuracy of 91.50 was achieved using LSTM.

In [8], a dataset was created by combining two CSV files: "factcheck.csv" from GitHub and "fake_or_real_news.csv" from Kaggle. NLP techniques such as count-vectorizer, TF-IDF, and hashing were applied. Various algorithms were employed, including Decision Tree, Random Forest, KNN, and Logistic Regression. The highest accuracy, 71%, was achieved by implementing Logistic Regression with TF-IDF vectorizer as NLP technique.

In [9], seven distinct datasets were used for the study. Machine Learning and Deep Learning models were used. The ML models included Naive Bayes and the Passive Aggressive Classifier. For ML, the data was TF-IDF vectorized, while for DL, tokenization was used as the Natural Language Processing (NLP) method. Except for one dataset where the results of the passive-aggressive and naive Bayes classifiers were similar, it was found that DNN outperformed both.

In [10], three different datasets were used – Liar [14], Fake or Real News, Combined Corpus. Authors of [10] carried out study on machine learning models, deep learning model and some pre-trained model. For traditional machine learning models, they used SVM, LR (logistic regression), decision tree and KNN. In deep learning models, they have evaluated six models for fake news detection. In traditional learning approach, Naive Bayes shows the best accuracy on all the three datasets. In the three datasets, there were three distinct winners among the six conventional deep learning models.

The authors of [11] employed a convolutional neural network model, a bi-directional long short-term memory networks model, a logistic regression classifier, and a support vector machine classifier. They found that CNN was the best performing models among all. This study was carried on LIAR dataset [14].

After getting data from several standard machine learning models, the authors of [12] used model prediction vectors from these models to arrive at a definitive classification result, which was either "real" or "fake." For a group of comparably successful models, an ensemble approach was utilized to counteract the drawbacks of a single model. At last, they validated the usernames, and by applying heuristics, they merged the results, after augmenting the heuristic result, accuracy of more than 95% was achieved.

In [13], advantage of multimodal models over unimodal models were discussed. Instead of using just a image, text associated with that image was also considered for detection of news. In this work, they looked at the role of tweet text and images for two issues related to the identification of conspiracies and fake news. They combined various CNN features for images with BERT features for text to achieve this.

As it is clear from above works that there is limited emphasis on comparing NLP algorithms in existing method, this study aims to fill this gap by conducting a systematic comparison of BoW and TF-IDF, while also evaluating their effectiveness in combination with various machine learning models for fake news detection. The results of this study can advance the field of fake news detection by improving knowledge of the function of natural language processing (NLP) techniques and how they affect the overall performance of fake news detection systems.

III. DATASET

In this study, we used the Liar [14] dataset to identify fake news. This dataset comprises political news statements authored by various journalists.

3.1. DESCRIPTION OF DATA

The dataset contains 14 columns, detailed description of the dataset content is tabulated below.

LIAR [14] dataset has 12,788 rows in all.

Column No.	Description
1	Statement's ID
2	label (i.e., True, or false)
3	statement (actual statement which our model will decide whether it is true or false)
4	Subject of news
5	speaker (the person who has given that statement)
6	Job title of speaker
7	state information
8	Concerned party (political party with which speaker is associated)
9-13	total credit history count, including the current statement
14	context

3.2. DATA PREPROCESSING

Given that all the political statements are in textual form, the first step involves converting them into numerical data. We did this by utilizing Natural Language Processing (NLP). Before applying NLP techniques, data cleaning is essential, which includes: Removing stop words, eliminating common words such as 'a,' 'an,' and 'the', Eliminating extra white spaces, removing punctuation marks, null values were removed.

After cleaning the data, we applied tokenization and lemmatization, Tokenization and lemmatization are two essential natural language processing (NLP) techniques used to process and analyse text data. They serve different purposes, but both are crucial for various NLP tasks, such as text analysis, information retrieval, and machine learning. Tokenization [15] is the process of breaking a text into individual units called tokens. Tokens are typically words, phrases, or symbols, and the purpose of tokenization is to split text into manageable pieces for further analysis. Lemmatization [15] is the process of reducing a word to its base or dictionary form, known as the lemma. Lemmatization aims to classify word forms that are derived or inflected together, so that they can be treated as once. This is useful for tasks like text classification and information retrieval, where you want to understand the core meaning of words. The Bag of Words (BoW) model is a basic method in natural language processing that represents text data as a vector of word counts. The two NLP algorithms that we are using are BoW and TF-IDF [15]. By ignoring word structure and order and concentrating only on the frequency and presence of specific words, it simplifies text. BoW is frequently utilized in tasks involving text classification and information retrieval. Word importance and frequency are combined in the more sophisticated text representation technique known as Term Frequency-Inverse Document Frequency (TF-IDF).

TF measures word frequency within a document, while IDF evaluates the significance of a word in the entire corpus. The TF-IDF score reflects a word's importance, making it a valuable tool in tasks like document retrieval, text mining, and information retrieval.

IV. METHODOLGY

4.1. RESEARCH DESIGN

This study adopts a quantitative research design aimed at investigating the effectiveness of two Natural Language Processing (NLP) techniques, Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), in conjunction with traditional machine learning (ML) models. The research design involves two main cases: Case 1 uses BoW as the NLP technique, while Case 2 employs TF-IDF. Each case is evaluated using four ML algorithms: Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Random Forest.

4.2. DATA PREPROCESSING

The dataset used in this study consists of labeled data comprising four key features: label, party, speaker [19][20], and text (a combination of subject and statement). Prior to analysis, unnecessary columns were dropped, leaving only the essential features required for the experiment. Label encoding was applied to convert categorical variables into numerical values, ensuring compatibility with the ML algorithms.

4.3. ANALYTICAL APPROACH

4.3.1 Data Preprocessing

After data collection, the preprocessing phase involved transforming the text data using NLP techniques. In Case 1, BoW was applied to represent the text data, while in Case 2, TF-IDF was utilized. These processed features were then used as inputs for the ML algorithms.

4.3.2 Model Selection and Optimization

The ML models selected for evaluation were Naïve Bayes, Logistic Regression, SVM, and Random Forest. To optimize their performance, Grid Search, a hyperparameter tuning technique, was employed. Grid Search helped identify the most advantageous parameters for each algorithm, thereby enhancing their overall performance.

4.3.3 Evaluation Metrics

The performance of each ML model was assessed using standard evaluation metrics such as accuracy, precision, recall, and F1 score. Cross-validation techniques were applied to ensure robustness and generalizability of the results.

4.4. STATISTICAL ANALYSIS

Statistical analysis was conducted to compare the performance of BoW and TF-IDF with respect to each ML algorithm.

DOI : 10.35291/2454-9150.2024.0065

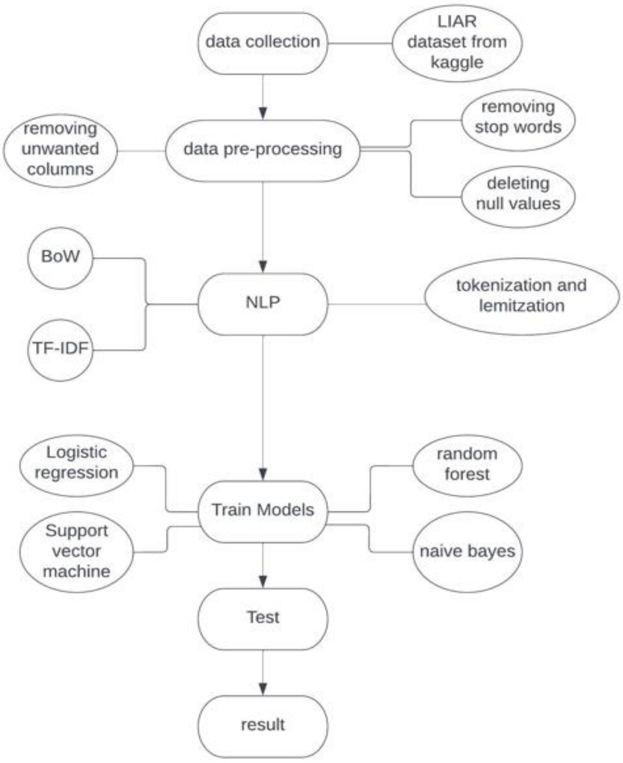

Fig 1: Flowchart – Proposed Model

4.5. MACHINE LEARNING MODELS

4.5.1 Naïve Bayes

We incorporated Naïve Bayes into our methodology, a probabilistic approach that relies on strong independence assumptions between features. Naïve Bayes essentially calculates the probability of our target variable given the occurrence of specific events or conditions. One key characteristic of Naïve Bayes is its simplicity and efficiency in modeling complex, high-dimensional data. It is particularly well-suited for text classification, spam filtering, and various real-world applications due to its ability to handle large feature spaces and make predictions based on the likelihood of certain events. This probabilistic approach is founded on Bayes' theorem.

4.2. Logistic Regression

Regression is based on the logistic function, which transforms input features into probabilities. These probabilities are used to predict binary outcomes, making it a valuable tool for tasks like binary classification and probability estimation.

Our model was fine-tuned using grid search. 3-fold cross-validation was used in our evaluation to determine the optimal hyperparameters and accuracy score.

4.3. Support vector machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for

regression and classification work. It decides which hyperplane best divides data points into discrete classes and maximizes the margin that separates them. Support Vector Machine (SVM) classification was used in our work, and it has been improved through grid search. The kernel functions selected are linear and rbf. 3-fold cross-validation was used to determine the optimal hyperparameters and the accuracy score that correlated with them. A detailed classification report was generated by evaluating the model on a validation dataset, and the model's performance was illustrated by a learning curve analysis for different sizes of training sets. By using parallel processing and a consistent random seed, reproducibility was preserved.

4.4. Random Forest

We used Random Forest classification, an effective ensemble machine learning method that synthesizes several decision trees to produce reliable and precise forecasts. Its ability to decrease overfitting and increase model stability makes it excellent in both classification and regression tasks. Our grid search strategy optimized hyperparameters like 'max_depth' and the splitting criterion ('gini' and 'entropy'). The model's resilience against overfitting and ability to handle complex data were leveraged. We assessed its performance via 3-fold cross-validation, resulting in the best hyperparameters and accuracy score.

V. RESULTS

Accuracies of all the models has been tabulated below

Algorithm	BoW	TF-IDF
Naïve Bayes	59.98	64.66
Logistic regression	66.62	67.29
SVM	62.84	65.32
Random forest	67.93	70.37

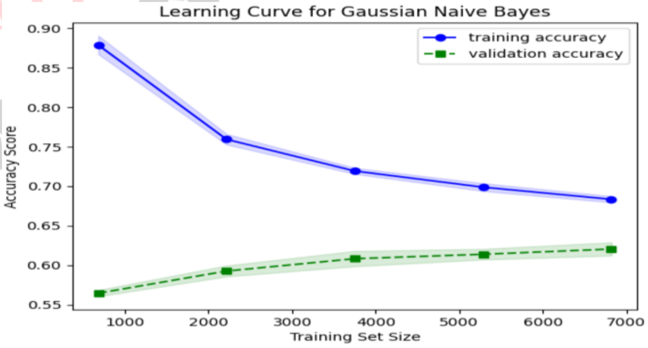

Fig 2: Naïve Bayes with BoW

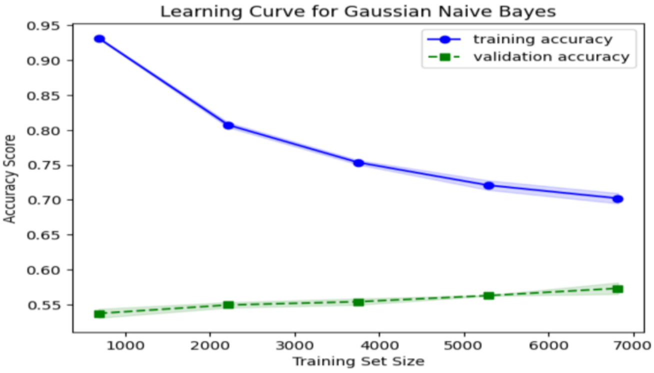

Fig 3: Naïve Bayes with TF-DIF

DOI : 10.35291/2454-9150.2024.0065

The graph in fig 2,3 shows that the training accuracy and validation accuracy of the Navies Bayes both increase as the size of the training set increases. This is expected as the training data increases but we can increase the training data more than 7000 as it is not yet reaching optimal state.

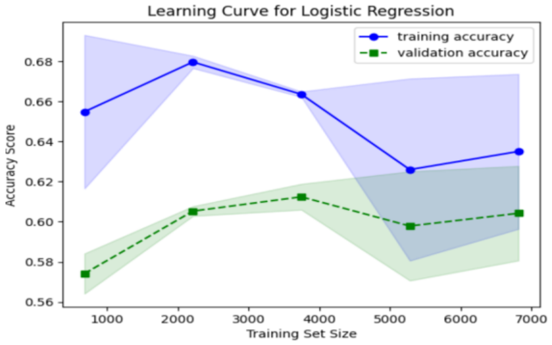

Fig 4: Logistic Regression with BoW

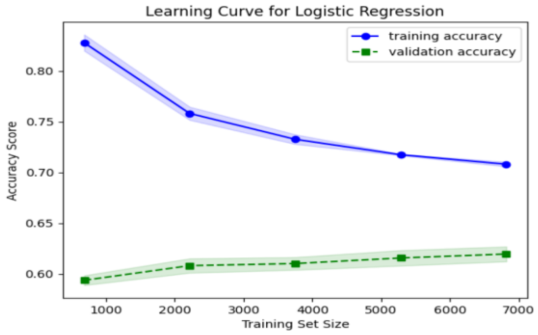

Fig 5: Logistic Regression with TF-DIF

In the graph fig 4 we can see due to underfitting the graph shows very big gap between training and validation accuracies but at around 4000 samples the graph getting plateau which is the indication that graph is overfitting if we increase more samples. Where as in TFIDF of logistic regression in fig 5 show the normal trend and samples size can be increases more.

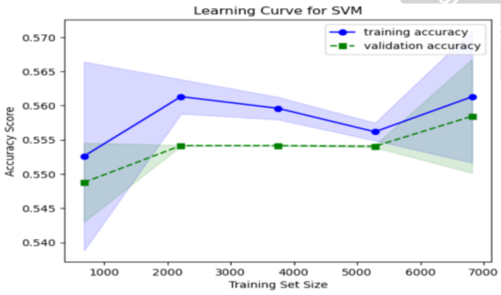

Fig 6: SVM with BoW

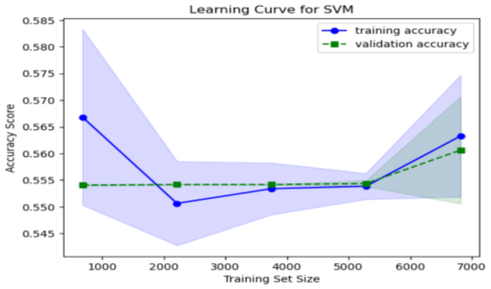

Fig 7: SVM with TF-DIF

The graph in fig 6,7 shows that the training accuracy and validation accuracy of the SVM both increase as the size of the training set increases. This is to be expected, as the SVM can learn more about the data as it has more data to train on. However, the graph also shows that the validation accuracy starts to plateau after a certain point. This suggests that the SVM is starting to overfit the training data, which means that it is learning the specific patterns of the training data too well and is not able to generalize well to new data. The optimal training set size would be the point at which the validation accuracy is highest without overfitting the training data. This is typically the point where the training and validation accuracy curves start to diverge. In this case, it looks like the optimal training set size is around 4,000 training examples.

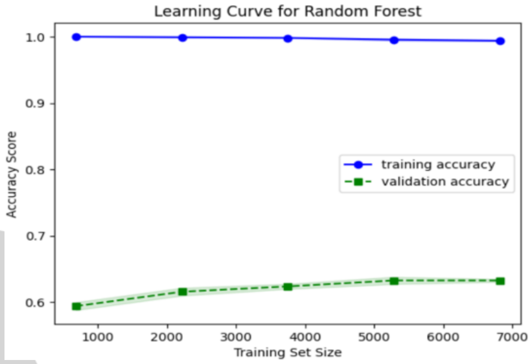

Fig 8: Random Forest with BoW

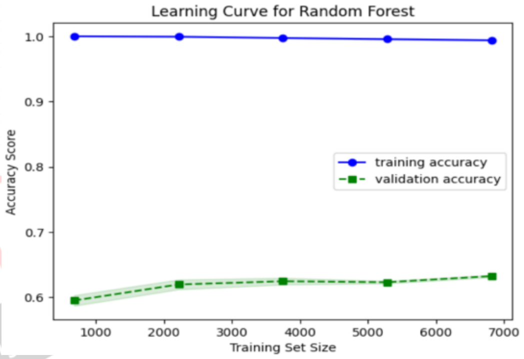

Fig 9: Random Forest with TF-DIF

In fig 8,9 random forest show the normal trends. In summary, all the graphs and various model shows the different accuracies but the normal trend shows the optimal training data size while be around 4000 samples as above that the graphs generally plateau and the accuracies either remain constant or it get overfitted in case of SVM TFIDF fig 7 where it shows the accuracies was close to each other.

VI. CONCLUSION

In conclusion, our study compared the effectiveness of Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) as Natural Language Processing (NLP) techniques in conjunction with various machine learning (ML) algorithms for fake news detection. We found that TF-IDF consistently outperformed BoW across all ML models, with Random Forest exhibiting the highest accuracy of 70.37%. While increasing training data improved performance initially, there was a tipping point where overfitting became a concern, particularly notable in Support Vector Machine (SVM). Random Forest demonstrated robustness with larger datasets, while TF-IDF with Logistic Regression showed potential for further improvement with increased data.

VII. FUTURE SCOPE

Moving forward, our research opens avenues for exploring deep learning architectures such as Recurrent Neural Networks (RNNs) or Transformer-based models for enhanced fake news detection. Additionally, comparing Word2Vec as an alternative NLP algorithm against BoW and TF-IDF could provide insights into optimal word representation techniques. Extending the scope beyond textual news to include audio and video formats in fake news detection would further enrich the understanding and applicability of our findings. These future directions aim to advance the field of fake news detection and contribute to more effective and comprehensive detection strategies.

VIII. REFERENCES

[1] Mayank Kumar Jain, Dinesh Gopalani, Yogesh Kumar Meena, Rajesh Kumar. *"Machine Learning based Fake News Detection using linguistic features and word vector features"* .2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).

[2] Anjali Jain, Avinash Shakya, Harsh Khatter, Amit Kumar Gupta. "*A smart system for fake news Detection Using Machine Learning*". 2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)

[3] Jasmine Shaikh, Rupali Patil. "*Fake News Detection using Machine Learning*". 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC).

[4] Shlok Gilda. "*Evaluating Machine Learning Algorithms for Fake News Detection*". 2017 IEEE 15th Student Conference on Research and Development (SCOReD).

[5] Wenlin Han, Varshil Mehta. "*Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation*". 2019 IEEE International Conference on Industrial Internet (ICII).

[6] Nihel Fatima Baarir, Abdelhamid Djeffal. "*Fake News detection Using Machine Learning*". 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH).

[7] Akash Dixit, Ishaan Kalbhor. "*Fake news Detection using Machine Learning*". 2022 IRJET e-ISSN: 2395-0056, p-ISSN: 2395-0072.

[8] Vanya Tiwari, Ruth G. Lennon, Thomas Dowling. "*Not Everything You Read Is True! Fake News Detection using Machine learning Algorithms*". 2020 IEEE.

[9] Rahul R Mandical, Mamatha N, Shivakumar N, Monica R, Krishna A N. "*Identification of Fake News Using Machine Learning*".

[10] Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, Anindya Iqbal. "*A Benchmark Study of Machine Learning Models for Online Fake News Detection*".

[11] William Yang Wang. *""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*". Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 422–426 Vancouver, Canada, July 30 - August 4, 2017.

[12] Sourya Dipta Das, Ayan Basak, Saikat Dutta. "*A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection*". arXiv:2101.03545v1 [cs.CL] 10 Jan 2021

[13] Gullal S. Cheema, Sherzod Hakimov, Eric Muller-Budack and Ralph Ewerth. "*On the Role of Images for Analyzing Claims in Social Media*".

[14] Kaggle Liar Dataset. Available at: https://www.kaggle.com/datasets/csmalarkodi/liar-fake-news-dataset

[15] Top NLP Algorithms. Available at: https://www.analyticssteps.com/blogs/top-nlp-algorithms

[16] Himangi Verma, Aditya Vidyarthi, Abhijit V. Chitre, Kirti H. Wanjale, M. Anusha, Ali Majrashi, Simon Karanja Hinga, "*Local Binary Patterns Based on Neighbor-Center Difference Image for Color Texture Classification with Machine Learning Techniques*", Wireless Communications and Mobile Computing, vol. 2022, Article ID 1191492, 11 pages, 2022. Available at: https://doi.org/10.1155/2022/1191492

[17] Arshpreet Kaur, Abhijit Chitre, Kirti Wanjale, Pankaj Kumar, Shahajan Miah, Arnold C. Alguno, "*Recognition of Protein Network for Bioinformatics Knowledge Analysis Using Support Vector Machine*", BioMed Research International, vol. 2022, Article ID 2273648, 11 pages, 2022. Available at: https://doi.org/10.1155/2022/2273648

[18] S. Maheshwari, How fake news goes viral: A case study, Nov. 2016. Available at: https://www.nytimes. com/2016/11/20/business/media/how- fake- news spreads.html

[19] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu. "User identity linkage across online social networks: A review". ACM SIGKDD Explorations Newsletter, vol. 18, no. 2, pp.5–17, 2017.

[20] Kai Shu, Suhang Wang, Huan Liu. "*Understanding User Profiles on Social Media for Fake News Detection*". 2018 IEEE Conference on Multimedia Information Processing and Retrieval.

DOI : 10.35291/2454-9150.2024.0065