ISSN No:-2456-2165

Fake News Classification using Machine Learning Techniques

Islam D. S. Aabdalla
PhD. Scholar of CSEJNTUH University
Hyderabad, India

Dr. D. Vasumathi
Professor of CSEJNTUH University
Hyderabad, India

Abstract:- Fake news exerts a pervasive and urgent influence, causing mental harm to readers. Differentiating between fake and genuine news is increasingly tricky, impacting countless lives. This proliferation of falsehoods spreads harm and misinformation and erodes trust in global information sources, affecting individuals, organizations, and nations. It requires immediate attention. To address this issue, we conducted a comprehensive study utilizing advanced techniques such as TF-IDF and feature engineering to detect fake news. We proposed Machine Learning Techniques (MLT), including Naïve Bayes (NB), Decision trees (DT), Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR) to classify news articles. Our studies involved analyzing word patterns from diverse news sources to identify unreliable news. We calculated the likelihood of an article being fake or genuine based on the extracted features and evaluated algorithm accuracy using a carefully crafted training dataset. The analysis revealed that the decision tree algorithm exhibited the highest accuracy, detecting fake news with an impressive 99.68% rate. While the remaining algorithms performed well, none surpassed the accuracy of the decision tree. This study highlights the immense potential of machine learning techniques in combating the pervasive menace of leaks. Our research presents a reliable and efficient method to identify and classify unreliable information, Safeguarding the integrity of news sources and protecting individuals and societies from the harmful effects of misinformation.

Keywords:- *Machine Learning, TF-IDF, Feature Extraction, Fake News Detection, social media.*

I. INTRODUCTION

Fake news is a term used to describe inaccurate or deceptive information that is presented as genuine news. This can encompass fabricated narratives, overstated or altered facts, and deliberately misleading content[1]. However, with new technologies, the internet has made it possible for people to access news from all over the world, at any time and on any device. The internet, primarily through social media and other media applications, has become the primary platform for spreading fake news. Despite the abundance of information available, the truth often needs to be clarified [2]. The purpose behind the spread of fake news is to manipulate the audience, whether for political or commercial gain [3]. In today's digital landscape, a vast amount of news is published across various media outlets, making it increasingly challenging to discern between accurate and false information [4]. Unreliable news creating for financial or political motives or to gain notoriety, using ideological narratives to deceive the receivers [5],[6]. This unreliable content, news manipulation, knowledge bubbles, and a lack of security on social platforms have become a pervasive disadvantage in our society.

Not only is unreliable news prevalent in traditional media, but it has also gained prominence in social forums, allowing it to spread quickly and extensively [2]. Clickbait, often with catchy headlines, is commonly used to attract readers' attention [7]. By clicking on these enticing titles, readers leading to poorly written articles with little relevance to the news they were expecting. Clickbait aims to drive more traffic to websites that rely on advertisements for revenue. An infamous example occurred during the 2016 presidential election, where Russian trolls used clickbait to sway public opinion away from Donald Trump toward Hillary Clinton. This instance illustrates the considerable influence that false information can exert on important matters. Social media platforms have evolved into environments where untrustworthy news, characterized by errors, informal language, and flawed grammar, proliferates[8]. The quest for improved credibility and accuracy has created an urgent need for techniques that help users make informed decisions [6].

Websites like Snopes and Politifact have emerged to fact-check news articles and uphold the truth. Research studies have also developed repositories to identify genuine and fraudulent internet sources [9]. In light of these discussions, categorizing unreliable news hinges on purpose and authenticity. Authenticity refers to false news containing inaccurate information. The second factor involves deliberately manipulating the news content to deceive the audience [10].

The main challenge lies in distinguishing between fake and real news [11]. Different social media platforms recognize false news through Extraction Features (FE), while traditional news societies rely on various factors, such as images and text, to identify and spot fake news. In terms of textual word-based sources, there are several aspects to consider:

It is essential to determine whether the article news carries the original content or just a part of it.

The authenticity of the news source needs to evaluate, knowing who published the news is crucial.

Another aspect to consider is the headline, which provides an in-detail news overview and aims to entice the audience. Additionally, the article news should accurately represent the content of the news. Researchers believe that detecting datasets and applying machine learning techniques can significantly contribute to quickly detecting unreliable news, both for the title and the article content [12]. However, categorizing article news poses a significant challenge due to analyzing text news from datasets, which involves processing many words, terms, and phrases, leading to computational limitations. Furthermore, redundant and extraneous features can harm the performance of classifiers. Feature engineering is crucial for enhancing performance. In this study, we bridge this gap by applying machine learning algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF). We also employ feature extraction techniques such as TF-IDF features, N-grams, and feature engineering.

The meaningful contributions of this paper are as follows:

- They are utilizing two datasets, removing unnecessary entities, eliminating duplicate and missing values, and merging them.
- After removing stop words and punctuation and converting text to lowercase, applying feature extraction techniques, such as TF-IDF, to the news articles, feature engineering is employed to enhance performance.
- It calculates the probabilities of each word and predicts whether it is fake or accurate based on these probabilities.
- To obtain the best results, we have implemented different algorithms for detecting fake news, including Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and SVM. We compare the performance of these algorithms with the previous approach. Notably, the decision-tree algorithm shows promising results in classifying junk news.

The remaining sections of this study follow the following structure: Section two provides a Literature Review, highlighting the related work on detecting unreliable news in the last three years. Section three presents the methodology framework for detecting fake news, focusing on models for predicting the news's authenticity. Section four presents the results and discussion, evaluating the obtained results. Finally, in section five, we conclude our study and provide recommendations for future work.

II. LITERATURE REVIW

This section provides an overview of relevant studies in the field. Additionally, numerous experiments have been conducted to detect the spread of fake news on social media using AI and ML.L. Sudhakar, M., and K. Kaliyamurthie [13] discussed the detection of fake news articles through ML algorithms. They identified several open problems that require further research. They proposed an LVQ (learning vector quantization) approach and achieved a precision output of 93.54%. The authors also suggested future research areas for the real-time identification of fake news in videos. Khan, J.Y., et al. [14] investigated the effectiveness of benchmarking ML models on various datasets for fake news detection. They analyzed the content and size of news articles and compared them with existing studies. The study aimed to assist the research community in selecting the most reliable technique for identifying fake news. The authors found that pre-trained BERT (Bidirectional Encoder Representations from Transformers)-based models performed well on small datasets.

Baydogan, C., and B. Alatas [15] proposed a framework based on ML models and NLP techniques to predict fake news from article content. They utilised different feature count vectors, word embedding, and TF-IDF (Term Frequency-Inverse Document Frequency) to generate feature vectors. The SVM (Support Vector Machine) linear classification algorithm achieved a precision of 0.94. B. Alatas and Ozbay [16] improved the detection of fake news articles by utilizing the FNC-1 dataset, which includes four categories of false news. They assessed modern techniques for fake news detection using ML algorithms and big data technologies. The authors employed a decentralized Spark cluster and stacked ensemble algorithms. By using N-gram, count vectorizers, and TF-IDF, they achieved a performance of 92.45% in detecting fake news.Amutha, R., and D.V. Kumar [17] presented a methodology for analyzing news information and distinguishing between real and fake news. They used a dataset consisting of Twitter microblog postings related to newsworthy topics. The study focused on supervised learning techniques such as SVM, decision trees, and Kappa statistics. The authors considered subsets of attributes, including text characteristics, social network features, and propagation-based attributes. SVM achieved high precision with 87% recall and 82% accuracy for real news and 84% precision with 89% recall and 87% accuracy for fake news.Kaur, P., and M. Edalati [18] analyzed and classified fake news using a dataset of approximately 40,000 news articles. They first created a list of stop-words to remove unnecessary words from the articles. Then, they applied CountVectorizer and TfidfVectorizer to generate feature vectors. They selected classification models such as Naive Bayes, Linear SVC, Logistic Regression, and Random Forest. Logistic regression achieved the highest performance, with 80% accuracy for fake news and 76% accuracy for reliable news.Meel, P., and D.K. Vishwakarma [19] focused on classifying movie opinions as positive or negative using ML algorithms. They analysed online movie reviews using opinion mining and text classification algorithms. Five ML algorithms, including DT-J48, SVM, NB, and KNN, were compared. SVM achieved the highest accuracy of 81.35% for sentiment classification. The authors also suggested extending the analysis to other datasets, such as those from Amazon or eBay.Aslam, N., et al. [20] discussed the reliability of news on the internet and proposed a fake news detection system. They collected posts from Facebook and used two classification techniques: a Boolean crowd-sourcing algorithm and logistic regression. Logistic regression achieved a high accuracy of 99% in predicting fake news posts.

This section presents an overview of relevant studies on detecting fake news. The approach adopted in this study aligns with the methods used by the previously mentioned authors. Moreover, various ML models are employed, feature extraction techniques are applied, including feature engineering. The study proceeds to compare different ML techniques and assess their effectiveness. When the results are compared to those of previous studies, this study shows exceptional performance.

III. METHODOLOGY

In this section, we present our proposed approach, which encompasses multiple stages, including using two datasets, feature extraction, feature engineering, ML classification, and addressing the challenge of detecting unreliable news.

The dataset comprises text news with attributes such as the headline, ID, and date, providing information on whether the news is real or fake. Figure 1 provides an overview of our approach, illustrating the process of detecting fake news on a combined dataset.

To begin, we merge two different datasets to create a corpus in the first step. The second step involves applying preprocessing techniques, including handling missing values, removing duplicate attributes, and eliminating unnecessary attributes in the fake news dataset. Furthermore, various preprocessing operations are performed on the news attributes, such as removing redundant words, converting text to lowercase, and implementing other necessary preprocessing steps. Subsequently, the dataset is divided into 80% for training purposes and 20% for testing, enabling further analysis.

In the third step, we concentrate on feature extraction methods to convert the textual data into numerical representations while utilizing feature engineering techniques to enhance accuracy. The fourth step details the ML models employed in this analysis as we explore various machine-learning algorithms for detecting fake news.

Finally, in the last step, we evaluate the performance of the models and compare them with other approaches, allowing us to assess their effectiveness.

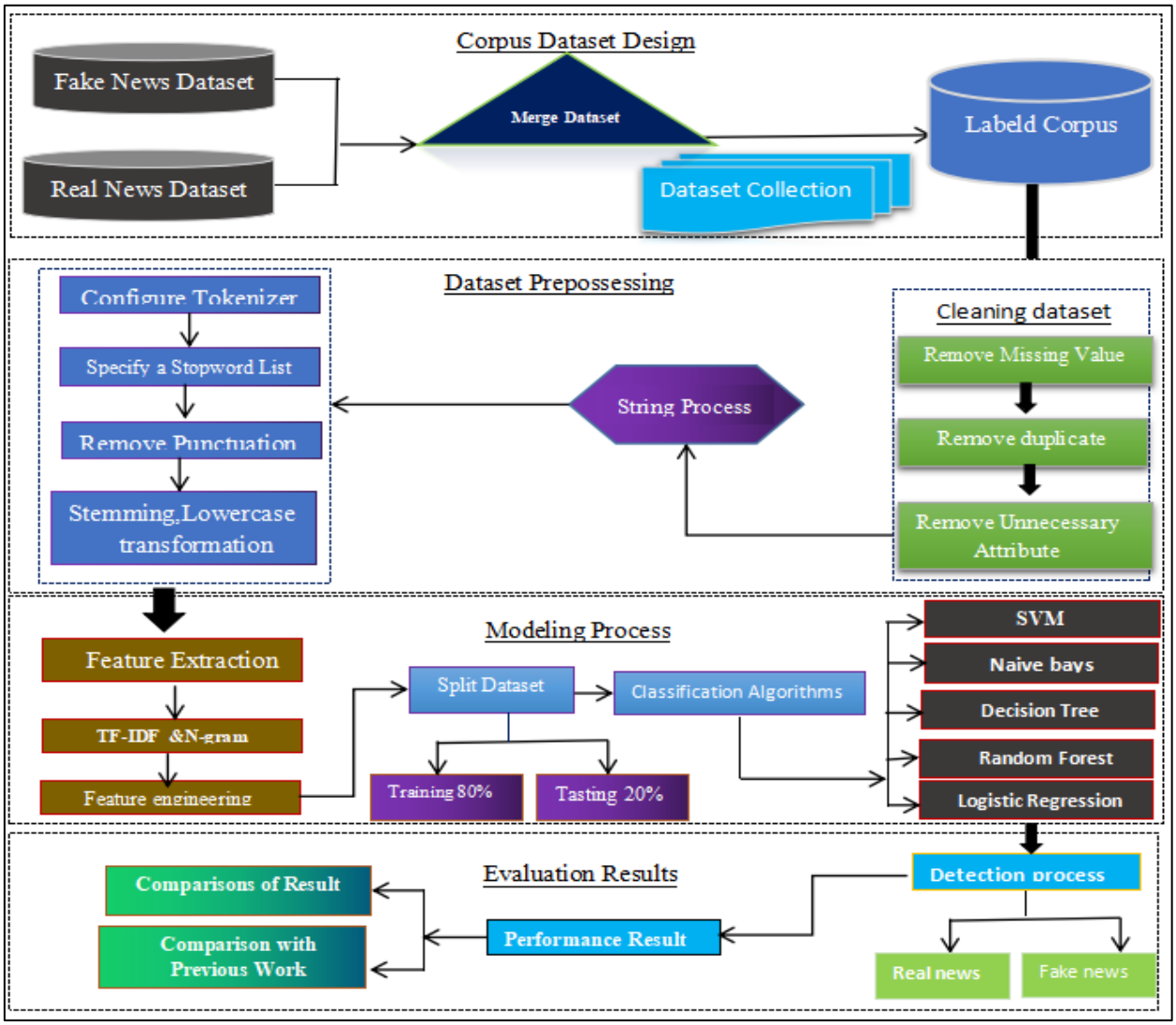

Fig. 1: Graphical Representation of Proposed System

ISSN No:-2456-2165

A. Dataset collection

This initial phase of the study focuses on the dataset. Table 1 presents a total of 23,481 headline-article pairs of fake news and 21,417 descriptions of actual news. Furthermore, Figure 2 provides a visual representation of the dataset, where the labeled class 0 represents fake news and 1 denotes real news. The dataset is accessible online on the Kaggle website [27].

Table 1: Discretion of Dataset Collection.

Attribute	Description
No	Unique ID for article article news.
Headline	The headline for articlenews.
Article news	Article news could be incomplete.
Class	A labeled of fake or real.
Author	How write article.
Date	Date of new make.

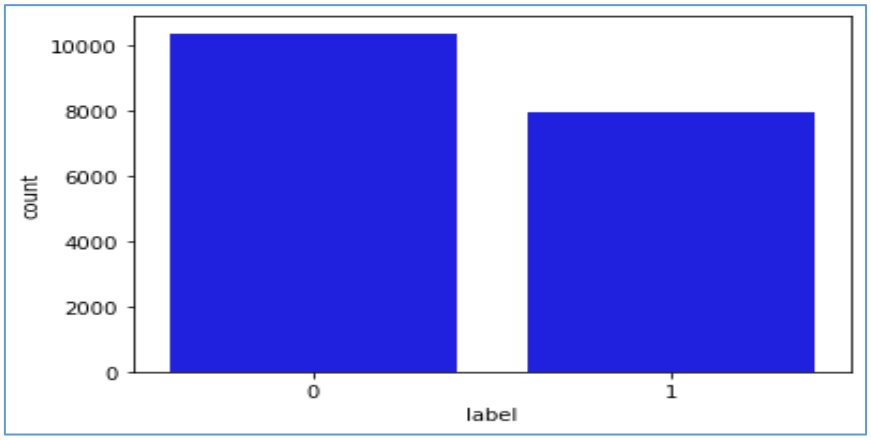

Fig. 2: Label Dataset fake and Real

Figure 3 provides an overview of the distribution of articles across different subjects. It includes 1,570 articles related to government news, 778 articles about medals in the East, 9,050 general articles, 783 articles on US news, 4,459 articles categorized as Left news, 11,272 articles labeled as P news, 10,145 articles covering world news, and 6,831 articles focusing on politics. These articles are sourced from reputable outlets such as the Washington Post, New York Times, CNN, etc. This study's findings validate the proposed model's effectiveness in identifying fake news articles by analyzing their text using machine learning algorithms. This approach dramatically streamlines the decision-making process.

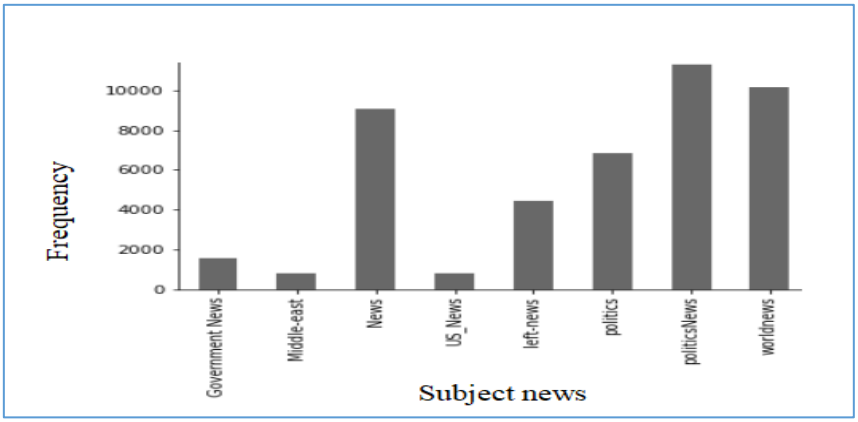

Fig. 3: Article News Per Subject

Moreover, Figure 4 and Figure 5 display word clouds that have been generated based on the identified fake and real news within the system, respectively. These word clouds visually represent the presence of multiple words associated with each category, offering an insight into the most frequently occurring words found in both fake and real news articles.

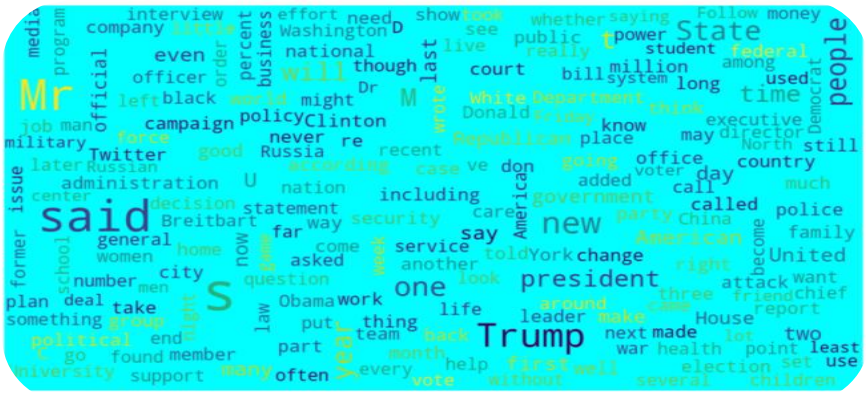

Fig. 4: Word Shadow for Unreliable News

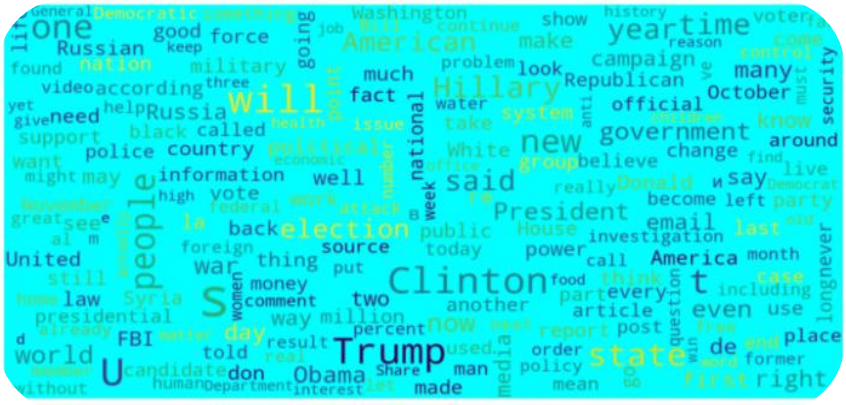

Fig. 5: Word Shadow for Reliable News

Table 2: Parameter Word cloud for fake and real dataset news

Target	Height	Max_Font_Size	Collocations	Ggenerate	Figsize	Width	Interpolation
ture	500	110	False	all_words	10,7	800	Bilinear

In Table 2, the most important properties for applying common words in real and fake news articles are presented. Furthermore, Figure 6 and Figure 7 provide visual representations of the distribution of these common words.

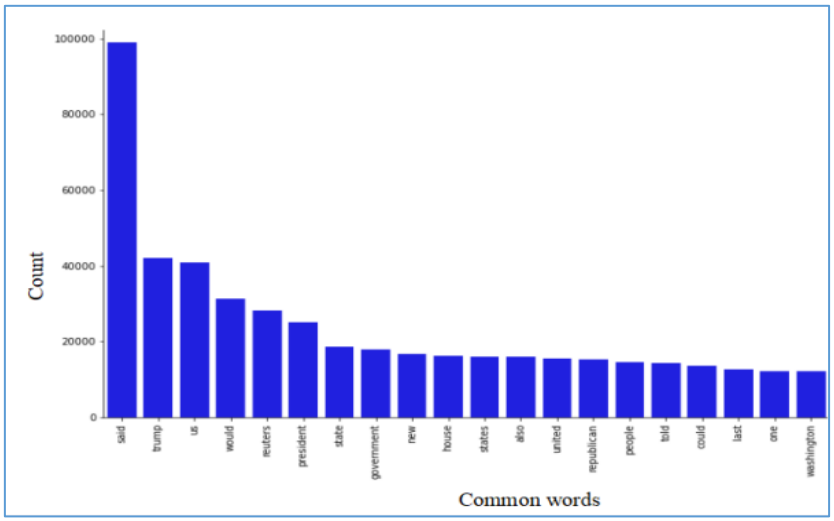

Fig. 6: Most Commonof real news

ISSN No:-2456-2165

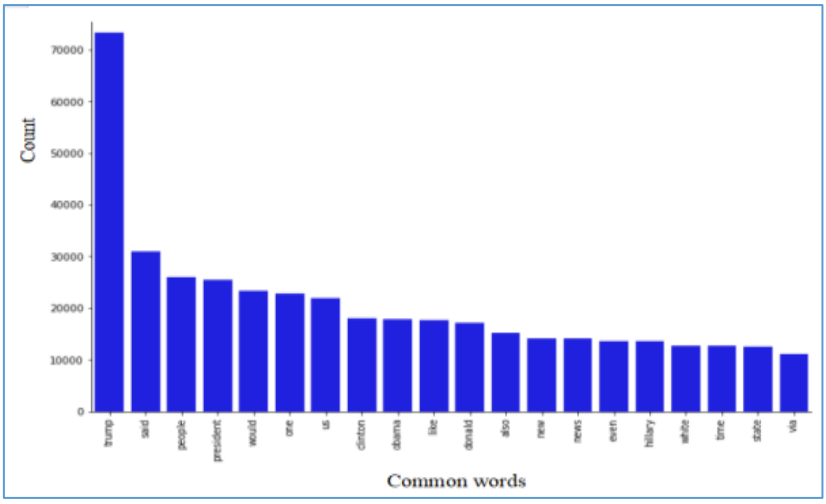

Fig. 7: Most Commonof fake news

Table 3: Common parameters

Parameters	
Rotation	Vertical
Figure size	12,8
Tocanization phrase	token_space.tokenize(all_words)
Data	Df_Frequency List(Frequency.Values())
Color	Red
Frequency	Nltk.FreqDist(token_phrase)
DataFrame(DF)	"Word": list(frequency.keys)
Df_Frequency.nlargest	N=Quantity columns = Frequency
Token_space	Tokenize.WhitespaceTokenizer

B. Prepossessing Dataset

Machine learning heavily relies on preprocessing to transform incomplete and inconsistent datasets into useful representations. Various text preprocessing techniques are applied to the dataset, including text transformation for stop word elimination, conversion to lowercase, stemming, tokenization, and utilization of models from the Keras library.

The dataset is then visualized using an N-gram term-based tokenizer, which segments the news based on the specified size of N. Specific preprocessing steps, such as tokenization, sentence segmentation, lowercase conversion, stop word removal, and punctuation deletion, are performed to reduce the dataset's volume by eliminating irrelevant details. These preprocessing steps are crucial in preparing the data for subsequent analysis. Data preprocessing plays a vital role in many supervised learning algorithms. The individual data preprocessing steps are as follows:

- Specify a stop words list and remove punctuation. Machine learning heavily relies on preprocessing to transform incomplete and inconsistent datasets into useful representations. Various text preprocessing techniques are applied to the dataset, including text transformation for stop word elimination, conversion to lowercase, stemming, tokenization, and utilization of models from the Keras library. The dataset is then visualized using an N-gram term-based tokenizer, which segments the news based on the specified size of N. Specific preprocessing steps, such as tokenization, sentence segmentation, lowercase conversion, stop word removal, and punctuation deletion, are performed to reduce the dataset's volume by eliminating irrelevant details. These preprocessing steps are crucial in preparing the data for subsequent analysis. Data preprocessing plays a vital role in many supervised learning algorithms. The individual data preprocessing steps are as follows:
- B. onfigure the tokenizer. Tokenization involves separating the news into units such as words or sentences. It facilitates text detection by converting the content into features using ML models [28]. After tokenizing the samples, the next step is to transform the tokens into a standardized form. Stemming is applied to convert phrases into their basic form, reducing the number of term types or labels in the data for faster and more efficient detection.

ISSN No:-2456-2165

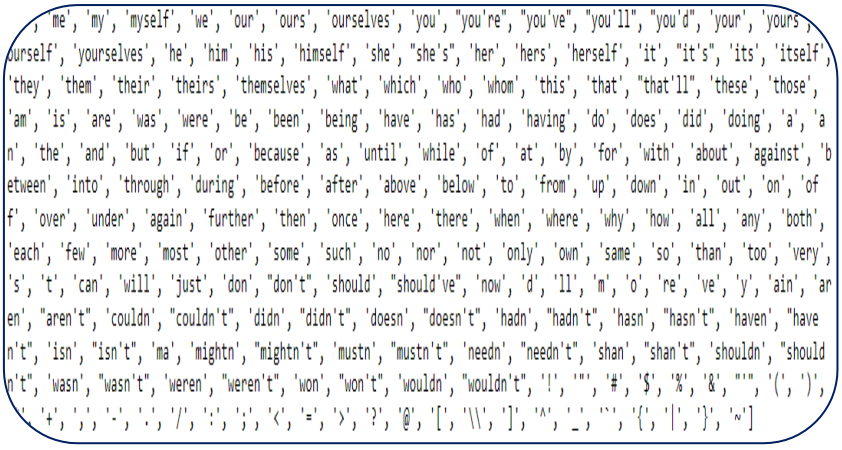

Fig. 8: Article News Tokenization

C. Lowercase transformation and stemming:

In this step, all terms in the dataset are transformed to lowercase to accommodate variations in capitalization. Moreover, stemming is applied using the NLTK's WordNet stemming implementation [8]. Conversely, the NLTK's Snowball stemming implementation [30] is utilized to reduce phrases to their stem forms. This rule-based approach aids in reducing the word corpus while preserving the meaningfulness of the words.

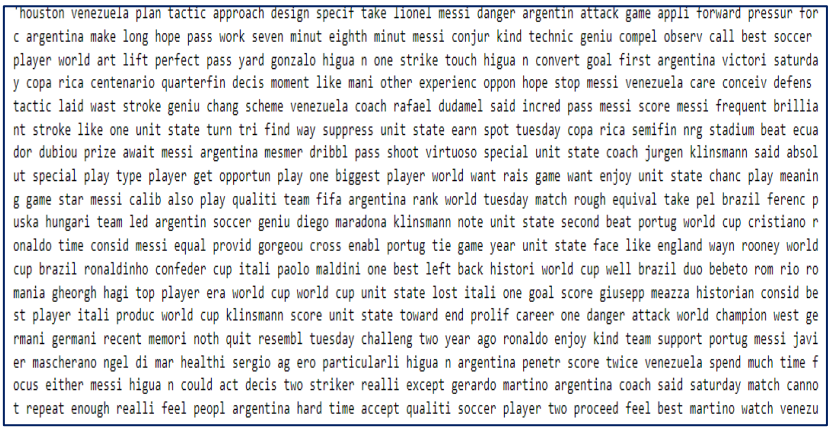

Fig. 9: Stemming and Convert to Lower Case Process

D. Feature Extraction (FE):

The main challenge in news categorization is dealing with high-dimensional data. The presence of numerous document terms, phrases, and words can lead to increased computational limitations in learning. Additionally, redundant and irrelevant features can hinder the interpretation of classifiers. Therefore, it is crucial to perform feature reduction and transform the text into numerical features that can be further processed while preserving the dataset [29].

The CountVectorizer of Words describes the occurrence of terms within news articles. It assigns a value of 1 if a term is present in the sentence and 0 if it is not. This creates a bag-of-words document matrix for each text document. N-grams are combinations of adjacent terms or phrases of length "n" that can be found in the original text [31].

TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used weighting metric in dataset analysis. It is a statistical measure that evaluates the importance of a phrase to a document in an article news. The reputation of a phrase increases with the number of occurrences within the document but is also influenced by its frequency in the entire corpus. The IDF (Inverse Document Frequency) It is computed by taking the logarithm of the ratio between the total number of documents in the corpus and the number of documents in which the specific phrase appears[3].

Term frequency (TF) is a method that uses the occurrence counts of terms in documents to determine the similarity between documents. Each vector is then normalized so that the sum of its elements corresponds to the total word count and represents the probability of a specific phrase existing in the documents [32].in the following equation:

$$TF = (\text{Number of occurrences of a term in a document}) \,/\, / \,(\text{Total number of terms in the document}) \quad (1)$$

$$IDF = \log D/(1 + DF) \quad (2)$$

Where:

- D is the total number of documents in the collection.
- DF is the number of documents containing the term.

[33]. For every word present in a dataset row, the value is non-zero, and if the word is not present, the value is zero. The TF-IDF of a token is calculated using the following two equations:

$$TF - IDF = TF * IDF \quad (3)$$

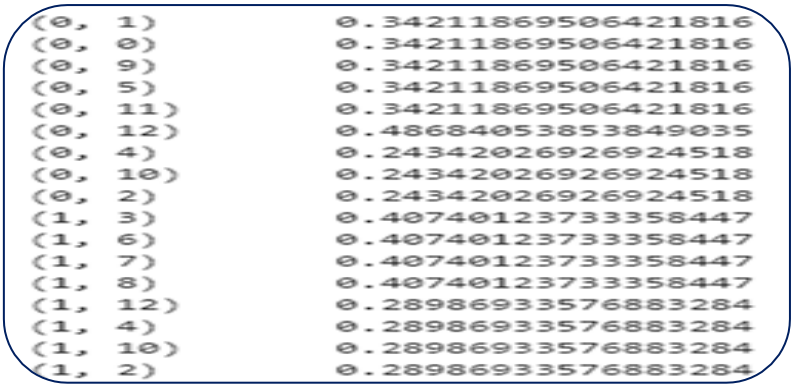

Fig. 10: Extract Article News

E. Feature engineering for fake news detection

Feature engineering (FET) is crucial for enhancing the performance of any machine learning algorithm, including its application to extract features from datasets. Transforming the raw dataset into feature data improves the quality of the model and enables achieving sufficient accuracy [30]. FET involves converting the original values and applying them during the feature engineering step. There are various techniques available for feature engineering, and sometimes it can be unclear which methods fall under the scope of FE and which do not [37].

F. Algorithms Used for Classification

Machine learning (ML) in real-time during the experimentation has a rapid impact on categorizing unreliable news. We use the following ML algorithms, such as Naïve Bayes (NB), decision tree (DT), Random Forest (RF), SVM, and Logistic regression (LR), to detect anomalies and analyze the effectiveness of our progressive algorithms.

➢ *Naïve Bayes (NB):*

The NB algorithm provides a probabilistic model-making technique. It computes the probability of each label variable's importance for conveyed input variable significances. By using dependent probabilities for an unexplored record, the model calculates the result of all target class weights and predicts the most likely outcome. NB is a classification algorithm that is probabilistic and supervised, originally developed by Thomas Bayes. It is easy to interpret and efficient for computation.

➢ *Decision Tree (DT):*

The DT algorithm partitions data into two or more subsets based on the similarity of samples. It is a recursive process that splits subsets and repeats the process until a stopping condition is satisfied. Each decision node tests the values of specific data functions, and each branch corresponds to a different test outcome. Decision trees are efficient for making classifiers and can handle both categorical and continuous variables.

➢ *Random Forest (RF):*

RF is a collection of tree predictors that depend on the same distribution for all trees. It uses a random vector to sample features independently in each tree. The prediction error for random forests converges, and they have the advantage of being robust against noise.

➢ *Support Vector Machine (SVM):*

SVM is a classification model that helps identify patterns in data for regression and classification. It creates learning processes from class training datasets and has a sound theoretical basis. SVM requires a relatively small number of samples compared to the dimensions of the data. It addresses the problem of discriminating between components of two classes using dimensional vectors.

➢ *Logistic Regression (LR):*

LR is a classification model used for predicting the outcome of a categorical dependent variable based on predictor features. It can handle numeric or categorical predictors and a categorical label. LR estimates discrete values and predicts the probability of an event occurring, with values between 0 and 1.

➢ *Evaluation Matrix:*

We use various assessment measures and evaluation metrics to analyze the efficiency of the model in detecting false news articles.

Accuracy: indicates the proportion of accurate predictions relative to the numeral of possible ones[8].

$$Acc. = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (4)$$

Recall: its point to the percentage of relevant measures retrieved from the whole numeral of relevant computed and instances[9].

$$Recall = \frac{True\ Positive}{True\ Positive + false\ negative} \quad (5)$$

F measure (F1 or F-score): harmonic mean of recall and precision [10]given by:

$$f - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

Precision indicates the percentage of actual test outcomes predicted accurately by dividing the numeral of correct predictions by the numeral of inaccurate ones[11].

$$precision = \frac{True\ Positive}{True\ Positive + False\ Postive} \quad (6)$$

This section presents the output or results of identifying fake news; the common word accurate and most common

ISSN No:-2456-2165

word fake contain the dataset, the classification models for real and fake news, classification models for opinions real and fake, and the evaluation of the results. This section has studied ML algorithms processing and analyzing datasets. Therefore, methods such as RF, NB, SVM, Support to determine and show which data are actual and which has been spreading fake over social media.

Table 4: Shows the Classification Report of the Proposed Model.

Methods	Accuracy (%)	Precision(%)	Recall.(%)	F1-score(%)
Naïve Bayes(NB)	99.55	99.72	99.42	99.56
Decision tree(DT)	99.68	99.69	99.71	99.69
Support Vector Machine(SVM)	94.86	96.75	93.36	95.02
Logistic regression(LR)	98.73	99.53	98.78	99.15
Random forest(RF)	99.03	99.37	98.78	99.07

The investigation results presented in Table 5 depict the performance of various classification models in identifying fake news using TF-IDF feature extraction and feature engineering techniques. Naïve Bayes demonstrates impressive results, boasting an accuracy of 99.55%, a precision of 99.72%, a recall of 99.42%, and an F1-score of 99.56%, underscoring its exceptional performance in fake news classification. The decision tree model stands out with the highest accuracy at 99.68% and exhibits commendable precision, recall, and F1 score. In contrast, SVM delivers satisfactory performance, achieving an accuracy of 94.86%, precision of 96.75%, recall of 93.36%, and an F1-score of 95.02%, albeit not reaching the levels attained by Naïve Bayes or the decision tree model. The logistic regression model performs reasonably well, securing an accuracy score of 98.73% and displaying good precision, recall, and F1-score (99.53%, 98.78%, and 99.15%, respectively), although there is potential for higher results. The random forest model showcases high accuracy (99.03%), precision (99.37%), recall (98.78%), and F1-score (99.07%), emphasizing its effectiveness in fake news classification. Figure 10 provides a visual representation of the comparison results, offering a comprehensive overview of each model's performance. These findings underscore the effectiveness of TF-IDF feature extraction and feature engineering techniques in enhancing classification accuracy and overall model performance. In summary, all models demonstrate strong capabilities in identifying fake news, with the decision tree model achieving the highest accuracy. Naïve Bayes, logistic regression, and random forest models also exhibit excellent performance, while SVM delivers satisfactory results. These outcomes emphasize the efficacy of the employed techniques in detecting fake news and provide valuable insights into the strengths of different classification models.

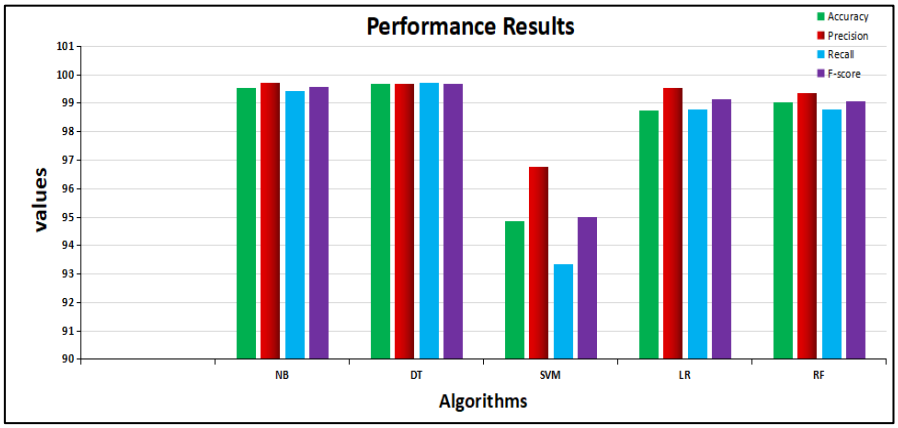

Fig. 11: Distribution of Classification Result

sequence follows: The distribution of the classification matrix of the ML algorithm is visualized in Figure 11. It depicts the number of instances for each class in the testing set. Accuracy was utilized to calculate the F1 score, precision, and recall as it pertains to the classification of the classes.

ISSN No:-2456-2165

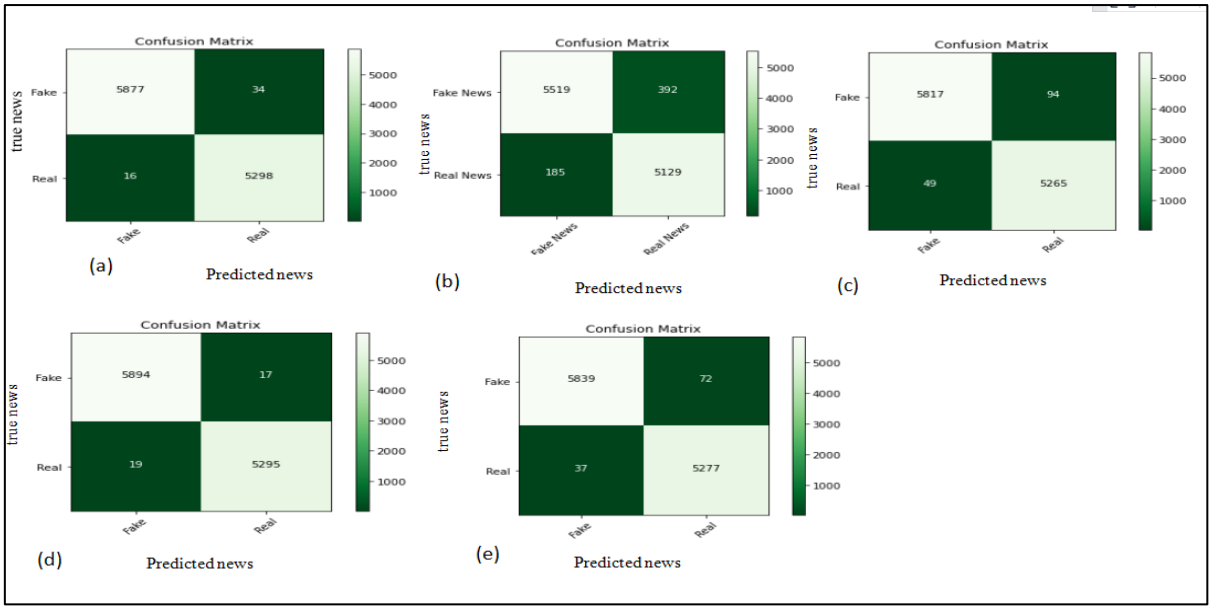

Fig. 12: Confusion matrix of final ML model (a)NB (b)DT (c) SVM (d)RF (e) LR

We also constructed the ML models, as shown in Figure 11. A confusion matrix is a table that provides an overview of the performance of supervised algorithms. The entries (A) NB, (B) DT, (C) SVM, (D) RF, and (E) LR indicate the models used, and they show that the models made some incorrect classifications. Among the models, the DT model achieved the highest accuracy of 99.68%, followed by NB with 99.55% and RF with 99.03%. Additionally, LR obtained a score of 98.73%, and SVM achieved a distribution of 94.86% as shown in Figure 7.

These results were part of the evaluation process, which included assessing accuracy, precision, recall, F1-score, and the confusion matrix to evaluate the model's performance.

Python was chosen for implementing the ML models due to its extensive libraries and high efficiency.

```
words_sentment("very good code but bad data",clf)

  FAKE      REAL

very      0.508011   0.491988
good      0.533461   0.466538
code      0.677308   0.322691
but       0.508011   0.491988
bad       0.559857   0.440142
data      0.495638   0.504361
------------------------------------------------
the sum   3.282287   2.717712
['the model prediction  ['FAKE
```

Fig. 13: Probability of the Fake News

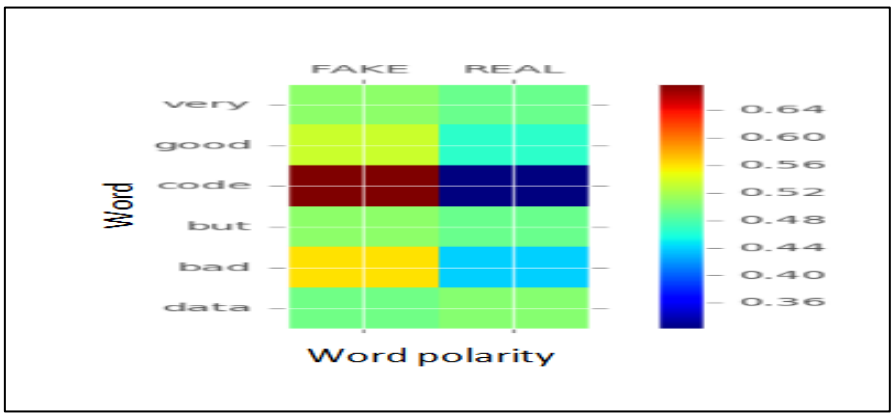

Fig. 14: Confusion Matrix of Each Words

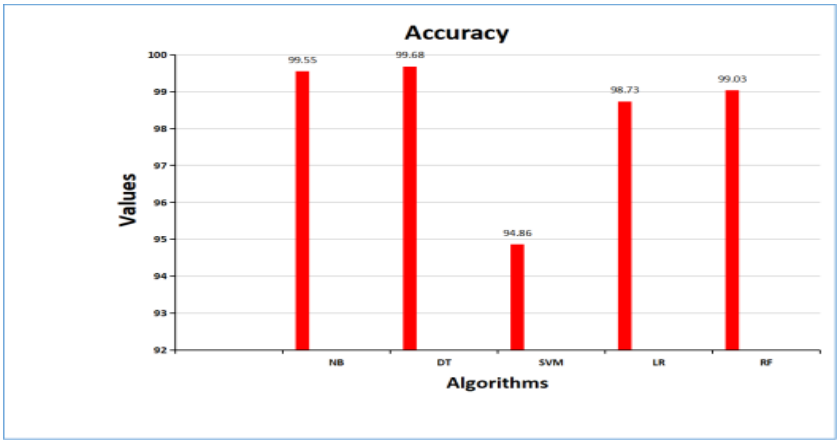

Fig. 15: Distribution of Accuracy Result

Following the feature extraction process depicted in Figures 13 and 14, the task was to identify unreliable or1141 real news articles by calculating the probability of being real or fake based on specific criteria. The models were trained using a dataset to estimate the probability values and analyze the dataset using three different methods. This allowed for determining which model was more accurate in classifying the news articles.

Table 6: Comparison-Based Classification Results with Previous Work.

Authors and References	Accuracy(%)
Faustini et,al.[3]	79.00
.Goswami et, al.[4]	85.86
Liu, Y. and Y.-F.B. Wu[2]	90.00
Altheneyan, A. and A. Alhadlaq[1]	92.45
Goldani, and S. Momtazi[7]	99.08
Our work	99.68

Table 6 provides a comparative analysis of our proposed model with previous studies in detecting unreliable news. The decision tree (DT) model in our current work achieved the highest accuracy of 99.68% for detecting fake article news, showcasing significant improvement.

IV. CONCLUSION AND FUTURE WORK

Many algorithms machine learning to detect fake news. However, it is crucial to select the model that achieves high accuracy on the datasets. This study focused on identifying fake news by utilizing feature extraction TF-IDF and feature engineering methods.

In conclusion, this study employed feature extraction using TF-IDF and feature engineering techniques to detect fake news. Several machine learning classification algorithms were applied and compared, including Random Forest, Naive Bayes (NB), Decision Tree (DT), SVM, and Logistic Regression. Our findings revealed that Decision Tree (DT) exhibited exceptional performance, achieving a remarkable classification accuracy of 99.68% in correctly identifying fake news, surpassing previous research results. For future endeavors, explore implementing deep learning algorithms further to enhance the development of real-time fake news detection techniques. Additionally, incorporating sentiment analysis into the detection process will contribute to better identifying and flagging fake news content. By integrating these advancements, we expect to significantly improve the accuracy and efficiency of detecting fake news.

REFERENCES

[1]. Romaguera, O.G., News (?) papers: A Typology of Fake News, 1880-1920. 2023.
[2]. Sarkar, S. and M. Nandan, A Comprehensive Approach to AI-Based Fake News Prediction in Digital Platforms by Applying Supervised Machine Learning Techniques, in Handbook of Research on Applications of AI, Digital Twin, and Internet of Things for Sustainable Development. 2023, IGI Global. p. 61-86.
[3]. Dumitru, E.-A., Testing children and adolescents' ability to identify fake news: a combined design of quasi-experiment and group discussions. Societies, 2020. 10(3): p. 71.
[4]. Fraga-Lamas, P. and T.M. Fernandez-Carames, Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality. IT Professional, 2020. 22(2): p. 53-59.
[5]. Khan, A., K. Brohman, and S. Addas, The anatomy of 'fake news': Studying false messages as digital objects. Journal of Information Technology, 2022. 37(2): p. 122-143.
[6]. Altheneyan, A. and A. Alhadlaq, Big Data ML-Based Fake News Detection Using Distributed Learning. IEEE Access, 2023. 11: p. 29447-29463.
[7]. Bruzzese, M., Fake news and information disorder: a journey through QAnon's conspiracy theory. 2021.
[8]. Hakak, S., et al., An ensemble machine learning approach through effective feature extraction to

classify fake news. Future Generation Computer Systems, 2021. 117: p. 47-58.
[9]. Singh, G. and K. Selva, A Comparative Study of Hybrid Machine Learning Approaches for Fake News Detection that Combine Multi-Stage Ensemble Learning and NLP-based Framework. 2023.
[10]. Liu, Y. and Y.-F.B. Wu, Fned: a deep network for fake news early detection on social media. ACM Transactions on Information Systems (TOIS), 2020. 38(3): p. 1-33.
[11]. Faustini, P.H.A. and T.F. Covoes, Fake news detection in multiple platforms and languages. Expert Systems with Applications, 2020. 158: p. 113503.
[12]. Weiss, S.M., et al., Text mining: predictive methods for analyzing unstructured information. 2010: Springer Science & Business Media.
[13]. Sudhakar, M. and K. Kaliyamurthie, Effective prediction of fake news using a learning vector quantization with hamming distance measure. Measurement: Sensors, 2023. 25: p. 100601.
[14]. Khan, J.Y., et al., A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 2021. 4: p. 100032.
[15]. Baydogan, C. and B. Alatas, Metaheuristic ant lion and moth flame optimization-based novel approach for automatic detection of hate speech in online social networks. IEEE Access, 2021. 9: p. 110047-110062.
[16]. Ozbay, F.A. and B. Alatas, Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: statistical mechanics and its applications, 2020. 540: p. 123174.
[17]. Amutha, R. and D.V. Kumar, Ensemble based Classification of Dynamic Rumor Detection in Social Networks for Green Communication. Journal of Green Engineering, 2021. 11(2): p. 1220-1243.
[18]. Kaur, P. and M. Edalati, Sentiment analysis on electricity twitter posts. arXiv preprint arXiv:2206.05042, 2022.
[19]. Meel, P. and D.K. Vishwakarma, Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. Expert Systems with Applications, 2020. 153: p. 112986.
[20]. Aslam, N., et al., Fake detect: A deep learning ensemble model for fake news detection. complexity, 2021. 2021: p. 1-8.
[21]. Kang, M., et al., A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on JD. com. Frontiers in Psychology, 2022. 13: p. 983060.
[22]. Zhang, X. and A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 2020. 57(2): p. 102025.
[23]. Kaliyar, R.K., A. Goswami, and P. Narang, DeepFakE: improving fake news detection using tensor decomposition-based deep neural network. The Journal of Supercomputing, 2021. 77: p. 1015-1037.
[24]. Vereshchaka, A., S. Cosimini, and W. Dong, Analyzing and distinguishing fake and real news to mitigate the problem of disinformation. Computational and Mathematical Organization Theory, 2020. 26: p. 350-364.
[25]. Di Franco, G. and M. Santurro, Machine learning, artificial neural networks and social research. Quality & quantity, 2021. 55(3): p. 1007-1025.
[26]. Goldani, M.H., R. Safabakhsh, and S. Momtazi, Convolutional neural network with margin loss for fake news detection. Information Processing & Management, 2021. 58(1): p. 102418.
[27]. Ahmad, I., et al., Fake news detection using machine learning ensemble methods. Complexity, 2020. 2020: p. 1-11.
[28]. Baarir, N.F. and A. Djeffal. Fake news detection using machine learning. in 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH). 2021. IEEE.
[29]. Wahab, O.A., Intrusion detection in the iot under data and concept drifts: Online deep learning approach. IEEE Internet of Things Journal, 2022. 9(20): p. 19706-19716.
[30]. Kaliyar, R.K., et al., FNDNet–a deep convolutional neural network for fake news detection. Cognitive Systems Research, 2020. 61: p. 32-44.
[31]. Perincheri, S., et al., An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Modern Pathology, 2021. 34(8): p. 1588-1595.
[32]. Rahman, M.S., F.B. Ashraf, and M.R. Kabir. An Efficient Deep Learning Technique for Bangla Fake News Detection. in 2022 25th International Conference on Computer and Information Technology (ICCIT). 2022. IEEE.
[33]. Lv, Z. and S. Xie, Artificial intelligence in the digital twins: State of the art, challenges, and future research topics. Digital Twin, 2022. 1(12): p. 12.
[34]. Shahid, W., et al., Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities. IEEE Transactions on Computational Social Systems, 2022.
[35]. Ragia Sultana, M.K.H., et al., An Effective Fake News Detection on Social Media and Online News Portal by Using Machine Learning.
[36]. Madani, M., H. Motameni, and H. Mohamadi, Fake news detection using deep learning integrating feature extraction, natural language processing, and statistical descriptors. Security and Privacy, 2022. 5(6): p. e264.
[37]. Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2021). Special issue on feature engineering editorial. Machine Learning, 1-12.