https://doi.org/10.33472/AFJBS.6.6.2024.9117-9132

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Elevating social media Fake News Detection through Feature Engineering Pre-processing using Ensemble Machine Learning Models

Parthiban.G[1], Dr. M. Germanaus Alex[2], Dr. S. John Peter[3]

[1]Research Scholar, Reg no: 21121282281017), Computer Applications, St. Xaviers College (Autonomous), Palayamkottai-627002 Manonmaniam Sundaranar University, Tirunelveli
[2]Assistant Professor, Department of Computer Science, Government College of Arts and Science, Nagercoil-629004
[3]Associate Professor, Department of Computer Science St. Xavier's College (Autonomous), Palayamkottai-627002

Email: [1]parthiguru2004@gmail.com, [2]mgalxus@yahoo.com, [3]jaypeeyes@rediffmail.com

Article Info

Volume 6, Issue 6, September 2024

Received: 27 June 2024

Accepted: 24 August 2024

Published: 12 September 2024

doi: 10.33472/AFJBS.6.6.2024.9117-9132

ABSTRACT:

In the contemporary field of information, combating Fake information on social media has become a formidable challenge. The Research paper employs on machine learning strategies, particularly focused on feature engineering, to enhance the accuracy of fake news detection. This involves selecting, transforming, and extracting relevant attributes from raw data to provide discriminative information to machine learning models. The pre-processing stage involves crucial steps such as identifying textual features, where the analysis of language patterns and sentiment helps discern Legitimate from Fake News. The methodology framework integrates ensemble machine learning models, such as Support Vector Machines, Random Forests, Artificial Neural Networks and Convolutional Neural Networks to effectively predict the authenticity of news articles or social media posts. The research explores an overall architecture that combines feature engineering models with neural networks for a comprehensive set of features. Experimental results show that ensemble classifiers, particularly combinations like Random Forest Classifier with Support Vector Machine (97.01), Artificial Neural Network (97.23) and with Convolutional Neural Network (98.21) significantly outperform individual classifiers. However, challenges such as dataset overfitting underscore the need for continual research and innovative approaches to address the evolving landscape of fake news on public platforms.

Keywords: Fake News Detection; social media; Feature Engineering; pre-processing; Ensemble Machine Learning Models;

1. Introduction

The abundance of social media in the trendy digital domain has changed the delivery of communication and oversaw problems such as the stretch of misinformation and fake news [1]. Solving the problem has become very important, leading investigators to explore progressive methods, and advanced mechanical engineering has appeared as a promising way to improve the exactness of models. look for false news. Fake information mentions the

dissemination of false or misleading information camouflaged as truth. The breadth and pace of the flow of information on social media have made it a breeding ground for the spread of false information. Researchers and data scientists are increasingly bending to sophisticated techniques, enclosing technology, to improve the cogency of fake broadcast observation techniques.

In the context of social media fake news detection, the prior generation of technology concentrated on extracting meaningful information from textual and contextual data and strengthening the footing of robust models. Analysis of the various languages of news reports and social media camps is an important arrow for the credibility of the announcement. Natural Language Processing techniques [2], such as sentiment analysis, were used to calculate the emotional tone of the text. In addition, semantic features such as word commonness and n-grams provided a lens to identify patterns that embody fake news. To correctly detect fake news, it is important to apprehend the context in which the knowledge is shared. Social network investigation is a vital tool for analyzing user relationships and identifying reporting patterns. Factors such as user engagement such as likes, shares and remarks were used as indicators of content credibility. Climate elements such as publication time and publication frequency elucidated the different patterns associated with fake news.

Exploring user behavior [3] related to sharing news scope added another wisdom into how the machine works. Factors such as publishing commonness, number of proponents, and past content-sharing patterns are useful needles. Modifications in user behavior, such as impulsive proliferation in performance, are red flags for the reach of fake news. Assessing the dependability of news reports is an important part of machine preprocessing. Factors affiliated with source reputation, historical exactness, and the inclusion of authoritative references were used as important indicators. Ensemble Machine Learning (ML)[4] techniques that analyze user behavior about trusted sources play an important role in evaluating source credibility. After pulling relevant features through pre-processing, machine learning models can organize news and social media posts as real or fake. Vigor was placed on well-known algorithms such as Support Vector Machines, Random Forests and Neural Networks. The foreword of pre-service technology strengthens these models by delivering valuable insights.

2. Related Works

Fake announcement refers to false or disingenuous information proposed as fact, often to mislead compilations. Social media scaffolds, with their large user grounds and instantaneous flow of information, have evolved into a forum for the spread of fake announcements. To sermons that problem, researchers and data scientists have twisted advanced techniques, including feature engineering, to increase the exactness of fake announcement detection prototypes.

Feature Engineering models [5] have mightily improved the exactness of fake news detection, but not without problems. Enemy attacks, as someone manipulates situations to hoodwink students, have become a major obstacle. Composers can create content that matches world content to exploit constituent set weaknesses.People use online platforms extensively to receive and distribute news, which helps both real and false stories spread widely. The spread of fake news throughout all social media platforms has a negative impact on society. One of the biggest challenges to effectively detecting false information on the social media platform is being unable to distinguish among the many types of incorrect information. Experts have improved their hunt for an answer by focusing on techniques to recognize false information. This research will make use of the set of data FNC-1, which has four different types for classifying fake news. The use of ensemble machine learning is used

to assess and compare the state-of-the-art techniques for identifying fake news. This study's technique used a centrally managed Catalyst cluster to build a model known as a stacked ensemble. After utilizing N-grams, Hashing TF-IDF, and count vectorizer for the extraction of features, **Altheneyan, A., &Alhadlaq, A. (2023)**[6]used the suggested stacked grouping model. The findings demonstrate that the recommended model outperforms the baseline method in terms of the accuracy of classification, with an F1 score of 92.45% as opposed to 83.10%. Comparing the suggested model to the current techniques, an extra 9.35% F1 score was attained.

Choudhury, D., &Acharjee, T. (2023)[7]presents a comparative analysis of several classifiers, including the SVM, Naive Bayes, a Random Forest, and Logistic Regression, for detecting fake news using various sets of data. With 61%, 97%, and 96% accuracy in the Liar, Fake Job Posting, and Fake News datasets, accordingly, SVM classifier has the highest accuracy. Once more, the fitness functions in our unique GA-based fake news detection algorithm are SVM, Naïve Bayes, Random Forest, and Logistic Regression. The LIAR dataset yielded 61% accuracy for both the SVM and LR classifiers in our suggested algorithm, while the fake job posting dataset produced the highest accuracy of 97% for SVM and RF.**Hanshal, O. A., et al. (2023)**[8] suggested a hybrid-enhanced model for deep learning to facilitate the identification of false news. In order to create artificially generated new fake news samples, the suggested model uses an automatic data augmentation technique known as Auxiliary Classifier Generative Adversarial Networks. It then combines recurrent and convolutional neural networks to effectively detect fake news. Using the Buzzfeed, FakeNewsNet, and FakeNewsChallenges datasets, the suggested model outperforms the state-of-the-art models with 93.87% accuracy, 10.39% recall, and 93.12% precision in identifying fake news.

In order to detect fake news from headlines and news descriptions, **Yadav, A. K., et al. (2023)** [9]examines a number of machine learning and deep learning techniques. This study creates an extensive set of data by merging two widely accessible datasets, fake and real news, and all Data, because there isn't a single, sizable, standard data set that can be utilized for fake news detection. The dataset contains 64,934 news articles with labels. Deep learning models, including CNN-BiLSTM, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Convolutional Neural Network-LSTM (CNN-LSTM), were used for classification. The CNN-BiLSTM model performed better on the established dataset when the Word2Vec word-embedding technique was used in conjunction with it. The accuracy, precision, recall, F1 measure, and AUC-ROC values were 0.975, 0.984, 0.970, 0.977, and 0.992, respectively. The suggested methodology performs better at identifying fake news than current cutting-edge techniques.While social media has grown in popularity as a news sharing and access tool, it has also contributed to the spread of false information, which presents significant risks. Concerns regarding the spread of false information are raised by the ease of dissemination and the continuous flow of information. In order to combat false news, timely news verification is essential. Nevertheless, South Asian languages have been overlooked in favor of English in the majority of research on fake information identification. **Roshan, R., Bhacho, I. A., & Zai, S. (2023)** [10]uses the hashing vectorizer and TF-IDF, two methods for text feature extraction, to analyze a dataset of Sindhi tweets. For analysis, a number of machine learning algorithms and sophisticated deep learning models, including Transformer BERT, were used.**Mohawesh, R., Maqsood, S., &Althebyan, Q. (2023)** [11] introduce an analytical method for detecting fake news that depends on relational variables that can be extracted directly from the text, such as organizations, sentiment, or realities. About 3.97%, 1.41%, 5.47%, 2.18%, and 2.88% better than the latest techniques for English to English, English to Hindi, English to Swahili, and English to Vietnamese language

examinations on the TALLIP fake news dataset were achieved by our model. To the best of our knowledge, this paper is the first to detect fake news in multiple languages using a capsule neural network.**Farhangian, F., Cruz, R. M., & Cavalcanti, G. D. (2023)** [12] carry out a comprehensive empirical investigation to assess various representation approaches and classification strategies according to their accuracy and computational expense. The outcomes of our experiments show that different dataset characteristics lead to different optimal feature extraction methods. Notably, transformer model-based context-dependent models consistently perform better. Furthermore, overall performance is enhanced by using transformer models as feature extraction techniques instead of just optimizing the network for the downstream task. The authorfinds through in-depth error analysis that, in order to achieve better generalization performance at a relatively low computational cost, a combination of feature representation techniques and classification algorithms—including classical ones—offers complementary aspect.

The explosion of fake news on social media platforms has created significant challenges for online news consumers. **Muduli, D., Sharma et al. (2023)**[13]proposed a customized CNN model called "Maithi-Net" to differentiate fake news from real news. The proposed model consists of five phases of convolution that can automatically learn the distinctive characteristics needed for recognizing fake news. The suggested model has been accurately validated using the ISOT and CGU-Maithili fake news data sets. Several evolution metrics, including accuracy, specificity, sensitivity, and the F1 score, are used to validate the model's effectiveness. With the CGU-Maithili dataset, the model's detection accuracy is 96.78%, while for the ISOT fake news dataset, it is 96.75%. The experimental findings in the field of fake news demonstrate notable improvements over earlier state-of-the-art findings.Identification and confirm the efficacy of our approach in classifying false material disseminated through social media.

False information propagated widely over online social networks (OSNs) has a number of detrimental effects. To identify fake news, a number of researchers developed various models utilizing deep learning (DL) and machine learning (ML) methods. The models' performance analysis revealed that their main shortcomings, such as inappropriate architectural design and inappropriate model suitability for various datasets, are the reasons for their low accuracy. A hybrid BERT-BiLSTM-CNN (BBC-FND) model is suggested by **Palani, B., & Elango, S. (2023)** [14]as a solution to these problems. An embedding layer, a feature representation layer, and a classification layer make up its three main layers. BERT is utilized in the first layer to extract the contextually-dependent features between newswords.In the feature extraction layer, a stacked BiLSTM and a multichannel CNN are used to generate different key features. While the latter extracts global semantic features from the contextual feature vector, the former extracts multiple features from the text and captures intricate local patterns in the spatial relations between words. To determine whether the news is fake or real, the concatenated features are fed into the FFN. The findings demonstrate that the BBC-FND model performs better than the SoTA approaches, with higher accuracy on four datasets (97.31 percent, 98.64 percent, 99.06 percent, and 98.26 percent) respectively.

Several deep learning model frameworks are used in the study to compare and identify misinformation in Chinese and English using various text feature selections. To make accurate or inaccurate predictions in the future, the model acquires knowledge of the textual attributes of both false and true information. The Gated Recurrent Unit (GRU) model, the Bidirectional Long and Short-Term Memory (BiLSTM) model, and the Long and Short-Term Memory (LSTM) model were chosen for the purpose of detecting fake news. The best detection result is produced by BiLSTM, which has an accuracy of 82% for Chinese texts and 94% for short- and long-sentence English texts, respectively.In order to recognize phony identities on social media, **Chen, M. Y., Lai, Y. W., & Lian, J. W. (2023)** [15] suggested

using recurrent neural networks. First, we use the Twitter API to extract data from social media platforms like Twitter.com. Based on the properties of the data, hybrid feature extraction has been carried out. It creates training rules connected to phony and authentic profiles created by humans. Using a policy-based approach, all bot entries are removed during the pre-processing and filtration stages. In order to produce stringent guidelines that enhance the classification precision, the dataset's training mainly concentrates on characteristics like the quantity of friends, total followers, tweets, retweets, and so on. Each profile is categorized by the Recurrent Neural Network (RNN) using training and testing modules. The goal of this work is to use machine learning to categorize people or robot entries based on the features that have been extracted.Following the initial training phase, capabilities are taken out of the dataset and applied to the term frequency using the classification technique. When it comes to identifying fraudulent accounts in a social media dataset that is unbalanced, the suggested work is quite successful. The social media dataset's fake and real identity classification is done with the highest level of accuracy possible by the system. With the various activation functions, it uses a recurrent neural network (RNN) to achieve good accuracy. As the number of cross-validation folds increases, the system's classification accuracy improves. We have tested both synthetic and real-time social media datasets for experiment analysis; we obtained approximately 96% accuracy on the real-time Twitter dataset and 98% accuracy on the synthetic social media datasets.

Shalini, A. K., Saxena, S., & Kumar, B. S. (2023) [16]proposed an incremental ensemble neural network system for detecting fake news that learns from fake news streams continuously and adjusts to changes. To improve overall performance, it uses pruning based on performance to remove underperforming classifiers. In order to preserve accuracy and robustness, the model also actively senses concept drift in real time and initiates adaptation strategies. Consistent news patterns are used for training and testing the models in two scenarios, allowing all ML and incremental models to perform consistently. In the second case, the study examines how news patterns change over time, taking into account concept drift brought on by momentous occasions like the US election. Performance degradation is a possibility for offline-trained methods, according to the analysis.Despite variations in the news pattern over time, the suggested model consistently performs well, achieving accuracy of 97.90% and 99.76% on two fake news datasets. The results show how fake news detection models' efficacy is affected by the news pattern's evolution. Consistent performance even in the presence of drift is indicated by the proposed model that was used for the experimentation.**Suryawanshi, S., Goswami, A., & Patil, P. (2023)** [17]examined machine learning techniques, including point-to-point analysis and dataset processing. Additionally, techniques like Random Forest, Naive Bayes, CNN, ANN, SVM, and all other potential methods that could assist in determining and delivering which data is real and which has been disseminating false information across social media globally have been compiled.

Disseminating false information regarding this conflict could have detrimental effects on both nations and their populations. We are therefore driven to develop a fake news detection system that is akin to those that are currently available in other industries, such as healthcare. **Babu, M. N. et al. (2023)** [18]assess the effectiveness of machine learning in this context by creating a dataset of real and fake tweets about the Russian-Ukrainian conflict. Using various dataset scenarios, the pre-trained BERT and five traditional machine learning algorithms, Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Naïve Bayes (NB)—are trained and assessed. The findings demonstrate that a system that can distinguish between authentic and fraudulent news about the Russian-Ukrainian conflict can be created.Compared to other learning algorithms, logistic regression and support vector machines yield prediction models that are roughly 76% accurate.**Darwish, O., et al. (2023)** [19]suggested a technique that leverages an ensemble of pre-trained

transformer models, each specifically optimized for the task of fake news detection, to more effectively identify fake news in languages with limited resources, like Hindi. We show that by overcoming the limitations of individual transformer models, the use of a transformer ensemble made up of XLM-RoberTa, mBERT, and ELECTRA is able to enhance the effectiveness of fake news detection in the language of Hindi.**Praseed, A., Rodrigues, J., &Thilagam, P. S. (2023)** [20]offers a useful framework that exclusively makes use of the news's textual elements. Using feature selection techniques, we determine the best-executing feature set that maximizes efficiency by evaluating multiple features for separating fake news from real news. Features for text representation were also investigated as a possible fix. In order to determine which model achieves the highest accuracy, tests were also conducted on the most popular machine learning and deep learning models. Our results show that high-accuracy fake news prediction can be achieved by combining text-based word vector sketches with semantic characteristics using ensemble methods. All other models were surpassed by Extreme Gradient Boosting (XGB), and linear support vector machines (SVM) produced results that were on section a similar level.

In order to detect COVID-19 fake news, this study investigates the efficacy of various machine learning algorithms and the optimization of pre-trained transformer-based models, such as COVID-Twitter-BERT (CT-BERT) and Bidirectional Encoder Representations from Transformers (BERT). **Chouliara, V., Koukaras, P., &Tjortjis, C. (2023)** [21] assess the effectiveness of various downstream neural network architectures with frozen or unfrozen parameters added to BERT and CT-BERT, such as CNN and BiGRU layers. Our tests on an actual COVID-19 fake news dataset show that adding BiGRU to the CT-BERT model improves performance significantly, achieving a state-of-the-art F1 score of 98%. These findings demonstrate the promise of cutting-edge machine learning models for the detection of fake news and have important ramifications for reducing the dissemination of COVID-19 disinformation.**Alghamdi, J., Lin, Y., & Luo, S. (2023) [22]** present a method for reliably detecting fake news. Information retrieval, the processing of natural languages, and machine learning make up the framework's three primary parts. The component parts of this study are the collection of data and the creation of machine learning models. Using a web crawler for information retrieval, we gathered the data from online news sources, and then [22] analyzed it using natural language processing techniques to extract pertinent information from web data. We selected a range of popular ML categorization methods for the comparison. [22] found that long short-term memory was the most effective model in the comparison investigation on the test set, and we developed an automated web application for spotting fake news on the internet.To ascertain whether a job posting is true or not, the suggested system from **Dev, D. G., & Bhatnagar, V. (2023)**[23]employs three distinct machine learning techniques: Support Vector Machines, Random Forests, and Naive Bayes Classifiers. Two methods were used to extract features from the data: Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Consequently, three models show respectable performance. Three separate machine-learning models are trained using various sample segments to create an ensemble model. The final predictions are determined by a simple majority vote among the three models. Consequently, three models show promise, with the Random Forest Model achieving an accuracy of more than 98.18%.

Few of the research gaps were predominantly witnessed in the ongoing research areas including the complexity of the dataset, imbalance of dataset in both positive and negative sides and also in overfitting of values during predictions. To address these issues, ongoing research from **Santhiya, P., et al. (2023)** [24] has concentrated on developing models that are vigorous against opponent attacks. Techniques such as malicious training, where the model is exposed to capitalize on data during training, can improve interpretation. To stay ahead of evolving competitor tactics, it is important to modernize technology and model

architectures.Another challenge lies in the dynamic nature of social media. The immediate evolution of user behavior, the formation of new features and changes in the algorithm database and the constant transformation of false information detection models. Both static feature extraction modes and dynamic extraction methods are being explored to ensure that models remain valid in the face of growing information environments

3. Materials And Methods

In today's communication world, the deploy of misinformation and false information on social media has become a prominent difficulty. To solve these problems, investigators and data scientists have used appliance learning strategies focused on feature engineering to enhance the accuracy of fake news detection. The technology involved selecting, transforming and extracting appropriate attributes from raw data to nourish discriminative information to machine learning models. In the circumstances of fake news observation in social media, the use of engine learning models is significant for extracting meaningful information from various parts of the data as shown in Fig.1.

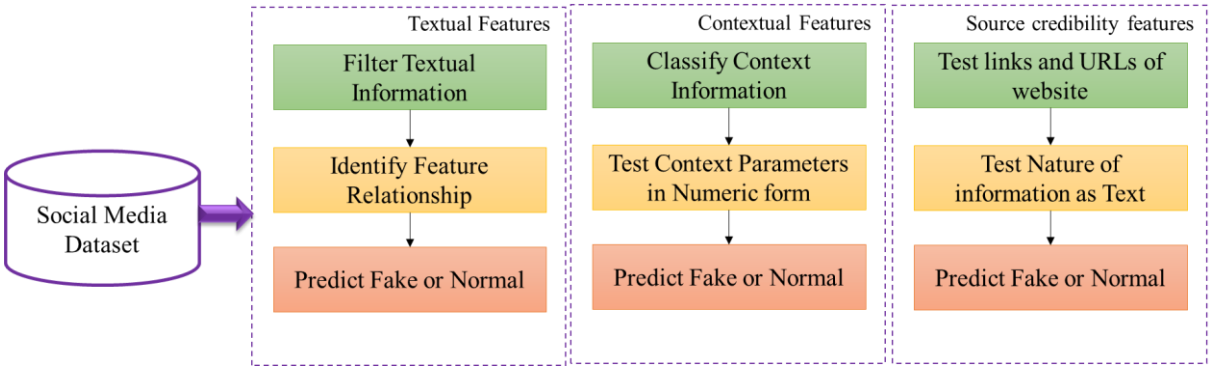

Fig.1. Various categories of features to be engineered in the pre-processing stage.

As shown in Fig.1., Processing is an important part of the engine learning process, the destination of which is to modify raw data into an arrangement that improves model enactment. In the field of fake news detection, the process involves identifying and extracting features that can discern between genuine and misleading information. Feature assignment is important because it affects the model's ability to generalize and make accurate prognoses.

3.1 Textual Features: The words and writing that are shared on social media are the most important part of the information that is posted. Feature engineering models work on finding important details in the text that can help identify patterns that show if the news is fake or not. Words, how often they are used, and groups of words are important in finding differences in language that might show something is not true.

False information often has strong emotions in it that try to control how people think, so sentiment analysis is a useful tool for dealing with this.

Furthermore, how often specific words are used and the way they appear together can show unique patterns connected to fake news. Identifying commonly used words in fake information helps the model tell the contrast among real and fake news.

3.2 Contextual Features: It's really important to understand the situation and surroundings in which information is shared on social media to spot fake news effectively. Feature engineering models use additional information to better understand how information is spread. Social media service is a way to study how people are connected and how information spreads between them.

The digit of relishes, portions, comments, and retweets shows how popular and believable the content is. Unusual changes, like a big increase in people paying attention, could mean fake news is spreading.

Adding the social network connections to the information helps the model tell if the news is from trustworthy sources or not, which helps to find fake news.

3.3 Source credibility features: Understanding how users behave is very important for finding fake news based on URLs and website links. Things like how often someone posts, how many people follow them, and what they have shared in the past says a huge about them. Changes in how someone acts online, like instantly attaching a lot or sharing a lot more than usual, could mean they are trying to harm others.

The variety of places where a user gets information from is another feature that can be created. Someone who shares content from many different places is persuasive someone who shares from only a few places. Checking if a news source is trustworthy is important for finding fake news. Feature engineering models combine features like the source's reputation, past accuracy, and the presence of reliable references in the content. Machine learning models can be taught to locating sample in how reliable sources are, using certain characteristics.

The history of when a news source has said sorry or fixed mistakes helps us know if it is trustworthy. Also, when an article has references or citations from experts, it is likely to be reliable. Collaborative filtering looks at how users behave and what sources they trust to help figure out if a source is trustworthy.

3.5 Ensemble Machine Learning Models: After finding important features and extracting them through feature engineering, machine learning models are used to determine if news articles or social media posts are real or fake. Many different computer models have been successful in this situation.

1. **Support Vector Machines (SVM):** SMV is a strong classification method used for telling the difference between two things, like real news and fake news.
2. **Random Forests:** Random Forests are a type of learning method that uses many decision trees together to make better predictions and to be more reliable.
3. **Artificial Neural Networks:** Neural Networks, like brain-inspired computer programs, are being used more and more to spot fake news. Neural networks can learn difficult patterns from data all by themselves, which helps them understand complex relationships in the feature set.

The Overall methodology framework for Fake news prediction is presented in Fig.2.

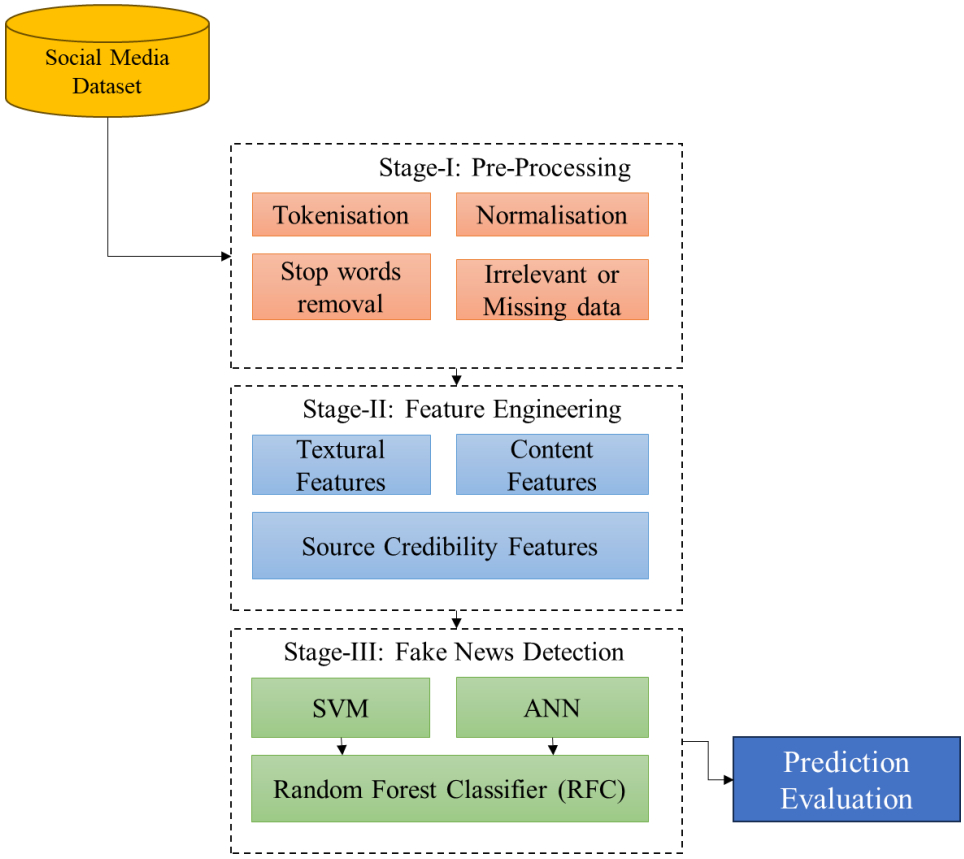

Fig.2. Overall architecture Framework of Fake News detection in social media

As shown in Fig.2., combining feature engineering models with neural networks customs creates the model perform better by giving it a complete and unique set of features. The type of machine learning program that choose depends on the details of the information and what they want our fake news detector to do.

IMPLEMENTATION AND EVALUATION

Fake news detection involves several steps, and pre-processing the data is a crucial part of building an effective machine learning model. The LIAR dataset [25], in particular, contains textual information about statements along with their labels indicating the truthfulness of those statements.The Structure of LIAR dataset is presented in Table.1.

Table.1. Structure of LIAR dataset with features and types.

S. No	Feature Name	Feature Type	Feature Category
1	the ID of the statement	Predictive	Textual
2	the label	Predictive	Textual
3	the statement	Predictive	Textual
4	the subject(s)	Predictive	Source Credible features
5	the speaker	Predictive	Source Credible features
6	the speaker's job title	Predictive	Source Credible features
7	the state info	Predictive	Source Credible features
8	the party affiliation	Predictive	Contextual
9	barely true counts	Predictive	Contextual
10	false counts	Predictive	Contextual
11	half true counts	Predictive	Contextual
12	mostly true counts	Predictive	Contextual

13	pants on fire counts	Predictive	Contextual
14	the context	Class	0 – Normal 1 - Fake

As shown in Table.1., the features of the dataset comprised of 14 different features with different columns as given in the subset.

$$Feature_{set(x)} = \{'ID','label','statement','subject','speaker','speaker\_title', 'state\_info','party\_affiliation','barely\_true\_counts','false\_counts', 'half\_true\_counts','mostly\_true\_counts','pants\_on\_fire\_counts','context'\}$$

The features of the dataset are loaded and the initial pre-processing is carried out to remove irrelevant or missing data and normalise the outcome data based on the test with URLs and website links. The set(x) has been classified based on the three categories of features including Textual, Source Credible features and Contextual as given in subsets below:

$$Feature_{set(x1)} = \{'ID','label','statement'\}$$
$$Feature_{set(x2)} = \{ 'subject', 'speaker', 'speaker'_{title},'state\_info'\}$$
$$Feature_{set(x3)} = \{ 'party\_affiliation','barely\_true\_counts','false\_counts', 'half\_true\_counts','mostly\_true\_counts','pants\_on\_fire\_counts','context'\}$$

All the Feature sets were independently engineered and pre-processed for further errors based on the Fake News Preprocessing Algorithm given in Table.2.

AlgorithmFakeNewsPreprocess FNP (Featureset(x1,x2,x3))
Begin Import Libraries of stopwords, stemmer; import nltk based stopwords; **# Preprocess text data** Compute stop_words:= initialise(stopwords.words('english')) Initialise stemmer asPorterStemmer() function Function preprocess_text(x1): Compute words := text.split(); Compute words := [stemmer.stem(word)] For word in words If word.lower() not in stop_words return ' '. concatenate(words) End For Update ['processed_statement'] asdf[‘statement’]. Apply(preprocess_text) **End**
End **FNP**

As given in Table.2., the text data has been split into different parts using stop words and based on the split of each and every text-based information, the stemming process is carried out to balance the datasets. Generally, the social media dataset is highly complex and huge resulting in overfitting of data at times. The dataset is also complex and unbalanced. Hence to balance the dataset, k-means clustering [26] could be employed to balance the outcome on all the sides of the dataset. This dataset is also stemmed with the stop words and after preprocessing process is completed, the text contents are merged again for normalisation process and missing data has been tested as well. Thus, at the end of the first stage of

prediction, the text, conceptual and source credible features were separately classified and normalised.

The Feature Engineering Process is the follow-up of the initial preprocessing stage where the features are selected in three different categories like Textual features, contextual features and Source Credible features respectively. These features were engineered based on the feature selection process of the high prediction features with the low prediction features of the dataset. The algorithm used in feature engineering process has been presented in Table.3.

Table.3. Feature Engineering process with Training of models for prediction

Algorithm Feature Engineering process with Model Training (FEMT){x1, x2, x3}
Begin Load the Trained_Model, Train_test_split; **# Split the data** X_train, X_test, y_train, y_test:= train_test_split(x1, test_size=0.2, random_state=42) **# Build and Train the Model** Import Training_Models for Feature_extraction Import Naïve_bayesExtractor_Model fromMultinomialNB Import pipeline for Feature_Engineering **# Build a pipeline with TF-IDF vectorizer and Multinomial Naïve Bayes classifier** Compute model := def make_pipeline(TfidfVectorizer(), MultinomialNB()) **# Train the model** model.fit(X_train, y_train) Display model (X_train, y_train) **End**
End **FEMT**

As shown in Table.3., the trained models were loaded for further training of the models and the split for the model has been claimed using the X_train, y_train inputs. The model fit is determined to verify the dataset has been balanced for better feature extraction process. To complete this process, the k-means clustering is applied to find the centroid of the prediction component. These features after arranged with the threshold of centroids are grouped as clusters. The clustered data is created as pipeline where the vectorised features are selected as MultinominalNB to train and predict the model using Naïve Bayesian classifier. The Naïve Bayesian Classifier is a supervised model that utilises standard feature extraction process thereby creates a goal directed feature engineering process. After completing the training and feature engineering process, the Prediction and Evaluation of Results process has been carried out as shown in Table.4.

AlgorithmFakeNewsPredict (FNP)()
Begin Declare accuracy_score, classification_report, confusion_matrix from performance metrics Declare validate_score_cross **# Make predictions** Compute y_pred:= model.predict(X_test); Apply Ensemble_Machine_Learning Models; Model.predict(rfc+svm); Model.predict(rfc+ann); Model.predict(rfc+cnn); **# Evaluate the model** Compute accuracy := accuracy_score(y_test, y_pred)

Compute classification_rep:= classification_report(y_test, y_pred) Compute conf_matrix:= confusion_matrix(y_test, y_pred) Display (accuracy) Display(classification_rep) Display(conf_matrix) **# Cross-Validation test** Compute cv_scores:= cross_val_score(model, process(x1,x2,x3), cv=5) Display Cross-validation scores Display Mean Cross-validation accuracy **End**
End FNP

As given in Table.4., the research uses accuracy, classification report, and confusion matrix to evaluate the model's performance on the test set.Cross-validation helps assess the model's generalization performance by splitting the dataset into multiple subsets and training/evaluating the model on different combinations of these subsets.These steps provide a basic framework for implementing fake news detection with the LIAR dataset using machine learning and pre-processing techniques. Remember that fine-tuning the model and experimenting with different preprocessing methods may be necessary for optimal results.

4. Results And Discussion

Creating tools in machine learning can help fix the problem of Fake news on public platform. By choosing and using important information from texts, user behavior, and source reliability, these models can better tell the contrast among real and false information.Text features show details in language, feelings, and word patterns, giving us an idea of what the content is about. Contextual features look at how people interact on social networks, how engaged they are, and how information spreads over time. This helps to realize the huge image of how information is shared. Ensemble models were tested in the prediction accuracy and performance. The overall architecture of ensemble Machine Learning Models applied to the results are furnished in Fig.3.

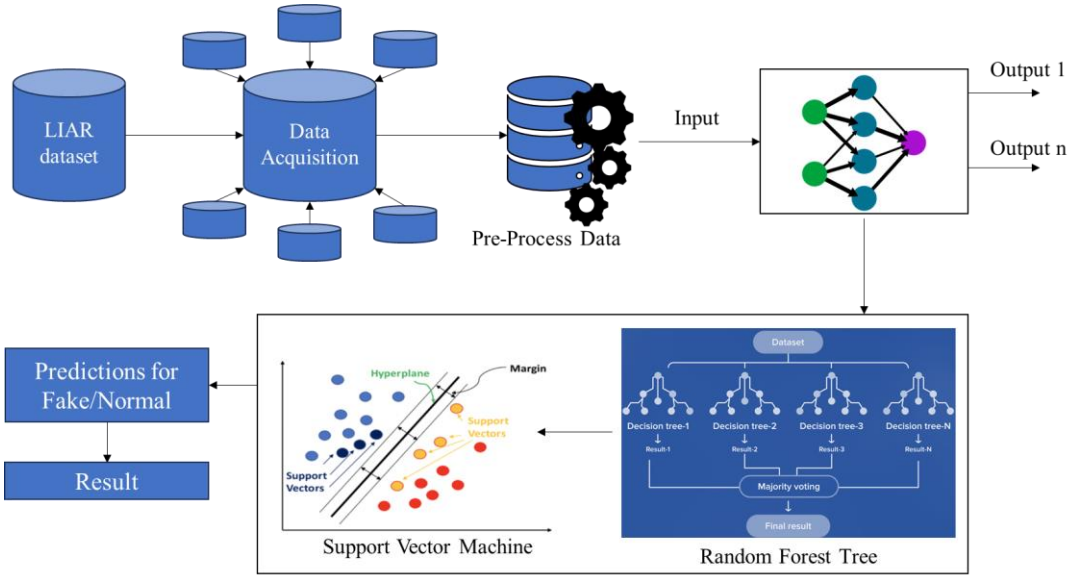

Fig.3. Overall architecture of ensemble Machine Learning Models using Random Forest Tree and Support Vector Machine

As given in Fig.3., the ensemble models like Random Forest Classifier (RFC) can be mingled with Support Vector Machine (SVM) to form an ensemble prediction model. The way people act and how believable a source is can say to a huge about how reliable the information is.Adding these features to machine learning models like SVM, Random Forests, and Neural Networks makes a strong system for finding fake news. However, problems like fake news and the always-changing social media sites need constant research and new ideas to create models that can handle these challenges. The dataset has been tested with different ML and DL classifiers and the efficiencies of performance measures achieved with regular classifiers of RFC, SVM, ANN and CNN as given in Table.4.

Table.4 Performance Measures of Individual Classifiers tested with LIAR dataset

ML & DL Classifiers	**Accuracy**	**Precision**	**Recall**	**F-Score**
RFC	61.76	68.57	66.76	68.78
SVM	74.54	77.65	74.87	78.76
ANN	78.65	79.53	76.65	77.67
CNN	87.55	88.77	87.66	89.98

From the Table.4., it was evident that the classifiers were not able to perform to the optimal level in normal prediction systems. Hence, the hyperparameters for ensemble models of ML and DL techniques were set for evaluation. The overall dataset was classified into 80% Training and 20% Testing sets respectively. The 10-fold Cross Validation is performed to test with 30 iterations to 40 iterations grouped ‘true’ and ‘mostly true’ as TRUE whereas the other outcomes of the cross validation fell into “FALSE”. The values are tested against the class values that contained six truth values {‘true’, ‘mostly-true’, ‘half-true’, ‘mostly-false’, ‘false’ and ‘pants-fire’}. The ensemble model first tests with the Random Forest Classifier (RFC) and then tests the outcomes of RFC with Artificial Neural Network (ANN), Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) to predict the outcomes of the proposed model as given in Table.5.

Table.5. Performance of Ensemble classifiers after cross validation predictions

Ensemble Classifiers	**Accuracy**	**Precision**	**Recall**	**Type-I Error**	**Type-II Error**	**F-Score**
RFC + SVM	97.01	97.34	97.48	0.78	0.24	97.87
RFC + ANN	97.23	97.41	97.44	0.76	0.15	97.46
RFC + CNN	98.21	98.33	98.36	0.73	0.09	98.55

Based on the outcome of the experiment suggested in Table.5., the classifiers performed well with outcomes in range of 97 to 99 outperforming the existing models without ensemble as given in Table.4. However, the ensemble models were tested with the outcomes of the confusion matrix. It was identified that the models couldn’t fit into the LIAR dataset and resulted in overfitting of dataset. Thus, dealing with the tricky problem of sham news on public platform needs advanced methods, and feature engineering preprocessing is a powerful way to enhance the perfection and strength of detection models. Researchers can create models to say the contrast among true and false information by using important details from the text, the context, how users behave, and how reliable the source is.

5. Conclusion

The research has profoundly presented an ensemble model for predicting the Fake News of the LIAR dataset with improved performance. The classifiers were also performing well with RFC combined with SVM, ANN and CNN classifiers. Although feature engineering preprocessing has been shown to enhance the validity of detecting fake news, there are still problems that need to be addressed. Intentionally changing things to fool models is a big problem. Working on making models stronger against attacks is highly essential making sure the news is highly reliable. Moreover, because public platforms are always changing, need to always update how detect things on them. The future works looks at ways to quickly analyze and extract important information from fake news to keep up with how it spreads.

6. References

1. Kondamudi, M. R., Sahoo, S. R., Chouhan, L., & Yadav, N. (2023). A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches. Journal of King Saud University-Computer and Information Sciences, 35(6), 101571.
2. Mahmud, M. A. I., Talukder, A. T., Sultana, A., Bhuiyan, K. I. A., Rahman, M. S., Pranto, T. H., & Rahman, R. M. (2023). Toward News Authenticity: Synthesizing Natural Language Processing and Human Expert Opinion to Evaluate News. IEEE Access, 11, 11405-11421.
3. Allein, L., Moens, M. F., & Perrotta, D. (2023). Preventing profiling for ethical fake news detection. Information Processing & Management, 60(2), 103206.
4. Alarfaj, F. K., & Khan, J. A. (2023). Deep Dive into Fake News Detection: Feature-Centric Classification with Ensemble and Deep Learning Methods. Algorithms, 16(11), 507.
5. Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content Based Fake News Detection with machine and deep learning: a systematic review. Neurocomputing.
6. Altheneyan, A., &Alhadlaq, A. (2023). Big data ML-based fake news detection using distributed learning. IEEE Access, 11, 29447-29463.
7. Choudhury, D., &Acharjee, T. (2023). A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers. Multimedia Tools and Applications, 82(6), 9029-9045.
8. Hanshal, O. A., Ucan, O. N., &Sanjalawe, Y. K. (2023). Hybrid deep learning model for automatic fake news detection. Applied Nanoscience, 13(4), 2957-2967.
9. Yadav, A. K., Kumar, S., Kumar, D., Kumar, L., Kumar, K., Maurya, S. K., ... & Yadav, D. (2023). Fake News Detection Using Hybrid Deep Learning Method. SN Computer Science, 4(6), 845.
10. Roshan, R., Bhacho, I. A., & Zai, S. (2023). Comparative Analysis of TF–IDF and Hashing Vectorizer for Fake News Detection in Sindhi: A Machine Learning and Deep Learning Approach. Engineering Proceedings, 46(1), 5.
11. Mohawesh, R., Maqsood, S., &Althebyan, Q. (2023). Multilingual deep learning framework for fake news detection using capsule neural network. Journal of Intelligent Information Systems, 1-17.
12. Farhangian, F., Cruz, R. M., & Cavalcanti, G. D. (2023). Fake news detection: Taxonomy and comparative study. Information Fusion, 102140.
13. Muduli, D., Sharma, S. K., Kumar, D., Singh, A., & Srivastav, S. K. (2023, June). Maithi-Net: A Customized Convolution Approach for Fake News Detection using

Maithili Language. In 2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3) (pp. 1-6). IEEE.

14. Palani, B., & Elango, S. (2023). BBC-FND: An ensemble of deep learning framework for textual fake news detection. Computers and Electrical Engineering, 110, 108866.
15. Chen, M. Y., Lai, Y. W., & Lian, J. W. (2023). Using deep learning models to detect fake news about COVID-19. ACM Transactions on Internet Technology, 23(2), 1-23.
16. Shalini, A. K., Saxena, S., & Kumar, B. S. (2023). Designing A Model for Fake News Detection in Social Media Using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 11(2s), 218-226.
17. Suryawanshi, S., Goswami, A., & Patil, P. (2023). FakeIDCA: Fake news detection with incremental deep learning based concept drift adaption. Multimedia Tools and Applications, 1-16.
18. Babu, M. N. V. V., Kumar, V. V., Vedavyas, T. K., Gampala, V., Chandra, S. D., &Thatavarthi, S. (2023, March). Machine Learning Approaches for Fake News Detection: A Review. In 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 132-137). IEEE.
19. Darwish, O., Tashtoush, Y., Maabreh, M., Al-essa, R., Aln'uman, R., Alqublan, A., ... &Elkhodr, M. (2023, March). Identifying Fake News in the Russian-Ukrainian Conflict Using Machine Learning. In International Conference on Advanced Information Networking and Applications (pp. 546-557). Cham: Springer International Publishing.
20. Praseed, A., Rodrigues, J., &Thilagam, P. S. (2023). Hindi fake news detection using transformer ensembles. Engineering Applications of Artificial Intelligence, 119, 105731.
21. Chouliara, V., Koukaras, P., &Tjortjis, C. (2023, June). Fake News Detection Utilizing Textual Cues. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 393-403). Cham: Springer Nature Switzerland.
22. Alghamdi, J., Lin, Y., & Luo, S. (2023). Towards COVID-19 fake news detection using transformer-based models. Knowledge-Based Systems, 274, 110642.
23. Dev, D. G., & Bhatnagar, V. (2024). Optimized Fake News Detection Model for Inspecting Data on the Cloud Using Machine Learning Techniques. In Integration of Cloud Computing with Emerging Technologies (pp. 213-226). CRC Press.
24. Santhiya, P., Kavitha, S., Aravindh, T., Archana, S., & Praveen, A. V. (2023, January). Fake News Detection Using Machine Learning. In 2023 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-8). IEEE.
25. William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.
26. Peng, L., Jian, S., Kan, Z., Qiao, L., & Li, D. (2024). Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. Information Processing & Management, 61(1), 103564.