DETECTION AND CATEGORIZATION OF FAKE NEWS WITH HYBRID APPROACH

|  |  |
| --- | --- |
| *Hemathi D*  *Department of Computer Science and Engineering Panimalar Engineering College*  Chennai [hemathimurthi@gmail.co](mailto:hemathimurthi@gmail.com) [m](mailto:hemathimurthi@gmail.com) | Dr T Jackulin  *Associate Professor*  *Department of Computer Science and Engineering Panimalar Engineering College*  Chennai [karthijackulin@gmail.com](mailto:karthijackulin@gmail.com) |
| *Ishani S*  *Department of Computer Science and Engineering Panimalar Engineering College*  Chennai [ishanisaravanan@gmail.com](mailto:ishanisaravanan@gmail.com) | *Madhumitha V*  *Department of Computer Science and Engineering Panimalar Engineering College*  Chennai [madhuvasanth1601@gmail](mailto:madhuvasanth1601@gmail.com)  [.com](mailto:madhuvasanth1601@gmail.com) |

***Abstract* - Fake news is becoming a growing concern, with the potential to influence public opinion and cause social chaos. As a result, there is an urgent need to develop effective fake news detection systems. The three main types of fake news detection approaches are content-based, social context- based, and knowledge-based. Content-based approaches analyse the news's content, such as text, images, and videos, to identify characteristics that are indicative of fake news. Text analysis, sentiment analysis, and credibility analysis are examples of such techniques. Knowledge-based approaches detect fake news by utilising external knowledge sources such as databases, ontologies, and external facts. The social context-based approach focuses on analysing the social context in which the news is being shared. This can include analysing the behaviour of the users sharing the news, the network structure of the users, and the diffusion patterns of information. Existing fake news detection solutions take one of three approaches: linguistic, knowledge-based, or social context-based. Here, we suggest a hybrid methodology for identifying fake news that includes both knowledge-based and content-based methods. We also divide the fake news into different categories according to its content. In comparison to other methods, our experimental results show that the proposed approach detects fake news with high accuracy.**

***Keywords -* Fake news, content-based, social context, knowledge-based, linguistic**

1. INTRODUCTION

The spread of fake news has become a big challenge in the digital environment in recent years. With the rise of social media and online news sources, distinguishing between authentic news and fraudulent information has become increasingly challenging. As a result, there is an increasing demand for dependable and effective systems capable of detecting and identifying fake news.

To fulfil this demand, a new method that blends the capability of natural language processing with knowledge-based reasoning has evolved. This hybrid linguistic and knowledge-

based approach leverages both linguistic features and background knowledge to identify and verify the accuracy of news articles

This system is built around a sophisticated linguistic analysis that evaluates the content and structure of news articles. The system analyses the language used in an article using machine learning algorithms to identify patterns that are indicative of fake news. The tone, sentiment, and style of the writing, as well as the presence of certain words and phrases commonly used in fake news articles, are examples of linguistic features.

In addition to linguistic analysis, includes a knowledge-based component that uses external sources of information to verify the accuracy of news articles. This includes factual information databases like news archives and government reports, as well as expert opinions and analyses from credible sources. The use of linguistic and knowledge-based analysis enables it to provide a more comprehensive and accurate evaluation of news articles. The system can quickly identify potentially fake news articles and provide a high level of confidence in its assessments by combining both approaches.

This hybrid linguistic and knowledge-based approach is a highly effective tool to combat against fake news. As the spread of false information poses a significant threat to the integrity of our information systems, systems like this will be critical in ensuring that accurate information is widely disseminated and trusted.

1. LITERATURE SURVEY

This section contains a review of the literature on existing fake news detection solutions. Approaches to detecting fake news can be classified into three types: linguistic-based, social context-based, and knowledge-based. N.Seddari.[1] proposes a novel approach for detecting fake news on social media using a hybrid linguistic and knowledge-based analysis and also demonstrated the effectiveness of approach through experiments on real-world datasets. G. Yildirim. [2] proposed a hybrid multi-thread metaheuristic approach for detecting fake news in social media, which combines the advantages of genetic algorithms and ant colony optimization . Thompson.R.C. [3] proposed a systematic review and meta- analysis methodology to quantitatively evaluate fake news

detection methods based on DL, ML, and ensemble. The meta- analysis was based on a database with nine variables related to these methods and data from 125 scientific articles. Effect sizes, heterogeneity, subgroup analysis, meta-regression analysis, and publication bias were all addressed for the included studies. This was due to the different sample sizes and approaches previously used in the methods. Deep learning, ensemble deep learning, ensemble machine learning, hybrid machine learning, and sentiment.Jarrahi.[4] proposed the CreditRank algorithm, which takes into account the network's publishers' activity history and credit rank. IT also demonstrated a novel sentence-level convolutional neural network (SLCNN) that can be used in text classification in general. SLCNN has the advantage of allowing other extra features to be combined at the sentence level. The SLCNN with CreditRank of publishers outperforms current methods. The proposed model was successful in detecting fake news with 99% accuracy. Ihsan Ali.[5] presented a study that focuses on the identification of false information and the application of cutting-edge detection methods at the news, user, and social levels. For identifying bogus news, it provided a taxonomy. The study looked into a number of advanced issues with false news detection systems. Moreover, a full analysis of techniques for detecting and recognising false news, including credibility-based, temporal- based, social context-based, and content-based techniques. The research examines various datasets used to spot fake news and suggests an algorithm as a last step.Raturi.S.[6]summarised various Machine Learning techniques used in detecting false news and the type of data we see on social media posts text, multimedia, or hyperlinks and there is notable achievement in detection of false news or fake posts with the use of various Machine learning approaches. Pandey.S.[7] proposed a work with the classifiers addressed the complex issue with an accuracy of 89.98% for KNN, 90.46% for Logistic Regression, 86.89% for Nave Bayes, 73.33% for Decision Tree, and 89.33% for SVM and also discovered that processing text for computation takes time when using Word2vec. Aside from that, it is simpler to execute the classifiers with a high accuracy report. Word2Vector is not usually recommended due to its high RAM and disc consumption, but it does provide semantic relations for processing data into vectors.Alwasel. [8] proposed a work focusing on detecting fake news in two stages: characterization and disclosure. In the first stage, social media is used to highlight the basic concepts and principles of fake news. During the discovery stage, existing methods for detecting fake news using various supervised learning algorithms are reviewed. POS textual analysis is a quantitative approach that is proposed to be added to Naive Bayes, SVM, and KNN methodologies. It depends on the addition of numerical statistical values as features and proposes to add in dataset the total words (tokens), total unique words (types), Type/Token Ratio (TTR), number of sentences, average sentence length (ASL), number of characters, average word length (AWL), nouns, prepositions, and adjectives, among other features. Sharma.[9] proposed a model was found to be a 65% accurate version of Logistic Regression. Logistic regression's performance was improved using grid search parameter optimisation, which resulted in an accuracy of 75%. There are 75% chances that it will be classified according to its true nature in this model.A.Jain. [10] proposed a review that the fake news detection using a machine learning and data mining approach. Future research will use a standard dataset to test the proposed model's ability to detect fake news and will use

a machine learning-based algorithm for detection and validation. The strategy will increase the performance metrics and create an effective algorithm.

1. PROPOSED METHODOLOGY

The steps used to achieve the goals of this study include the following:

1. Data collection
2. Data Preprocessing
3. Feature extraction
4. Modeling Creation with Random Forest
5. Hyper parameter Tuning

Figure 1 depicts the architecture diagram of the method.

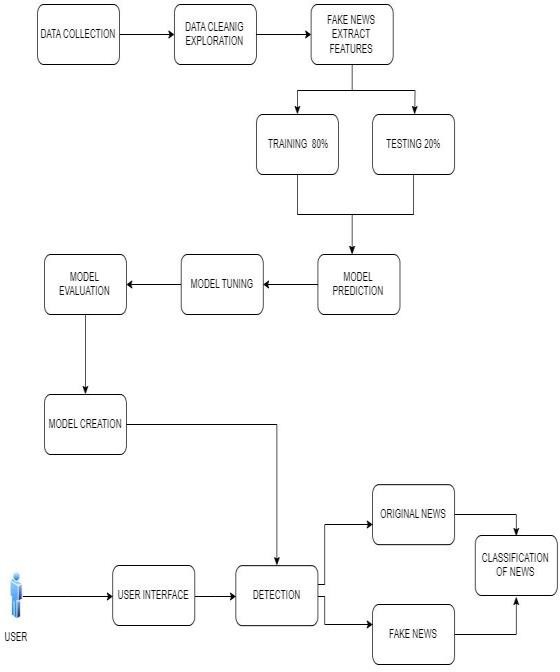


Fig. 1. Architecture Diagram

1. DATASET

Kaggle is a popular platform for hosting and sharing datasets, and it provides a wide range of datasets for different use cases, including fake news detection. While Kaggle datasets can be a good starting point for developing fake news detection models, it is important to carefully evaluate the quality and relevance of the data before using it for training and evaluation.

Some considerations when using Kaggle datasets for fake news detection include:

Data quality: Check the quality of the data and ensure that it is reliable, consistent, and accurate. Look for any missing values, inconsistencies, or errors in the data and address them appropriately.

Data bias: Evaluate the potential biases in the data and ensure that they do not affect the performance of the model. Look for any skewed distributions, underrepresented groups, or other biases that may affect the model's ability to generalize to data.

Data relevance: Ensure that the data is relevant to the problem being addressed and covers a diverse range of news articles and statements. Look for datasets that include both fake and real news articles, and cover different topics and domains.

Data size: Consider the size of the dataset and ensure that it is large enough to train a reliable model. Look for datasets that are sufficiently large and diverse, and can be used to train models with good generalization performance.

Data licensing: Check the licensing terms of the dataset and ensure that it is suitable for your intended use. Look for datasets with clear and permissive licensing terms that allow for commercial or non-commercial use.

1. PREPROCESSING
2. *Text Cleaning*

The first step is to clean the text data by removing any irrelevant information such as special characters, punctuation marks, and stop words. The text data can also be converted to lowercase to ensure consistency.

1. *Tokenization*

The text data is broken down into individual words or tokens. This step involves splitting the text into separate tokens, which can be further processed or analyzed.

1. *Stemming and Lemmatization*

Stemming and lemmatization are techniques used to reduce words to their root form. This step is important to reduce the dimensionality of the feature space and to remove the variations in the text data.Stemming is the method of decreasing words to their root form, whereas lemmatization is the method of decreasing words to their dictionary form.

1. *Feature Extraction*

The next step is to extract the relevant features from the text data.The feature extraction process entails transforming text data into a numerical format that can be used by the machine learning models. Common feature extraction techniques include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings.

1. *Removing Stop Words*

Stop words are commonly occurring words in a language that do not carry much meaning. Examples of stop words include "the", "and", "in", "is", etc. Removing stop words can reduce the dimensionality of the feature space and enhance the model's performance.

1. *Handling imbalanced datasets*

Fake news datasets are often imbalanced, with a larger number of real news instances than fake news instances. Handling imbalanced datasets is important to ensure that the model is not skewed against the majority class. Techniques

such as undersampling, oversampling, and SMOTE (Synthetic Minority Over-sampling Technique) can be used to balance the dataset.

1. PERFORMANCE METRICS
2. *Accuracy*

It measures the analysis of TP and TN to the total no. of test case data.

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 = (𝑇𝑃+𝑇𝑁) / (𝑇𝑃+𝑇𝑁+𝐹𝑃+𝐹𝑁)

1. *Precision*

It is the estimation analysis of true positive to the aggregate value of true positive and false positive rate. It is given in eqn.

Precision = (Number of correctly identified fake news articles) / (Number of articles identified as fake news)

1. *Recall*

It is the estimation analysis of true positive rate to the aggregate value of the true positive and false negative rate. It is given in eqn.

Recall =(Number of correctly identified fake news articles) / (Number of actual fake news articles)

1. F1 Score

The F1 score can be used as an evaluation metric for fake news detection models, which are typically binary classifiers that aim to identify whether a news article is fake or real.

In this case, the F1 score would be calculated by comparing the model's predictions to the true labels of a test set of news articles. The true labels would indicate whether each article is fake or real, and the model's predictions would indicate whether the model predicted each article to be fake or real.

A high F1 score for a fake news detection model would indicate that the model has a good balance of precision and recall in identifying fake news.

F1-Score = {2.Precision.Recall}/{Precision + Recall}

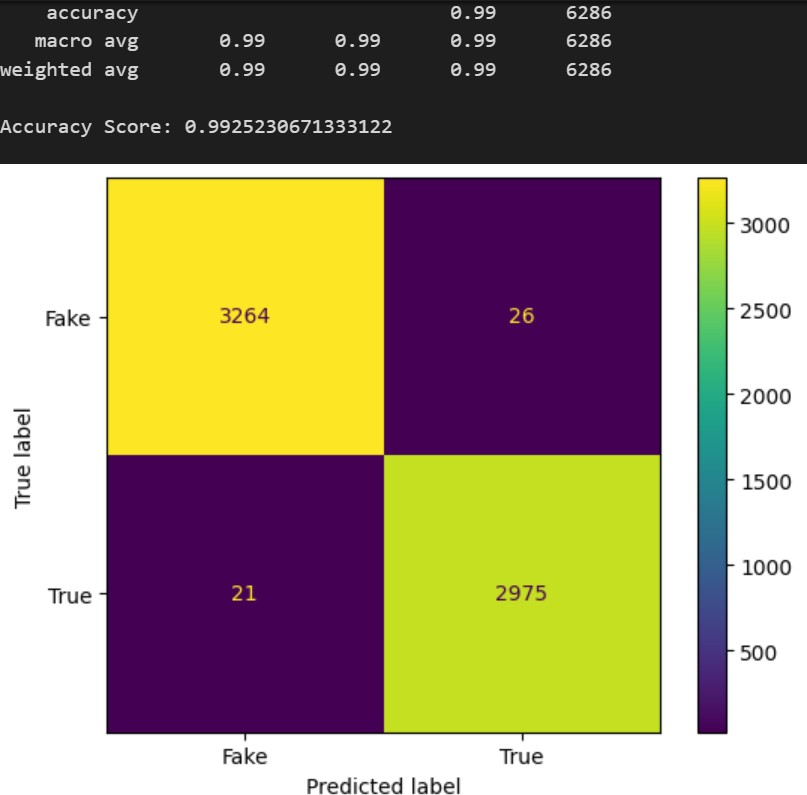
1. EXPERIMENTAL RESULTS

In this paper four classifiers have been implemented, and their performance in classifying the supplied article collection is compared. We used the Confusion matrix for this purpose. A confusion matrix shows how many misclassifications and accurate classifications the model made. The observed outcome in terms of confusion metrics.

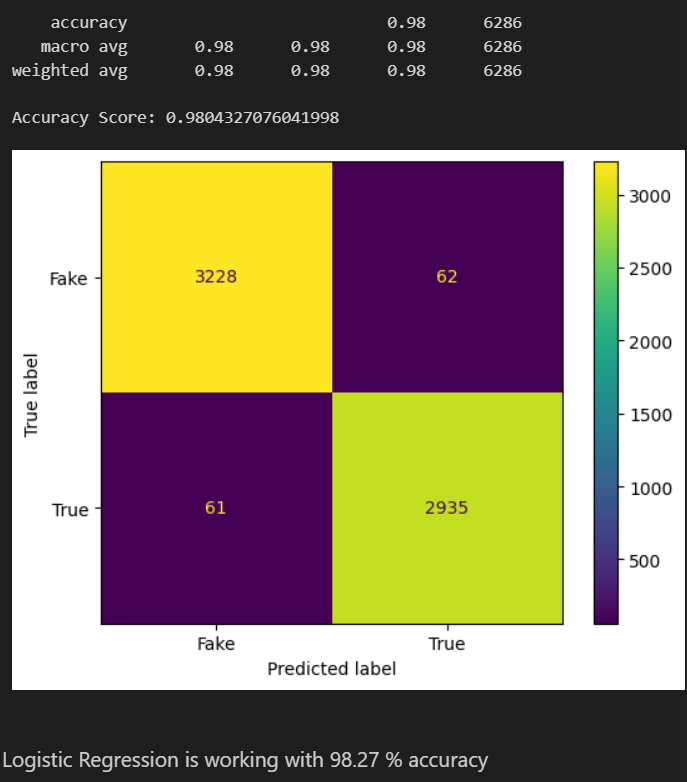
Given that the classifier categorised the bogus news as positive, there are four alternative parts, which are explained below:

* The items that have been accurately recognised as fraudulent are labelled in the top left area, which is referred to asositives that are true positive.
* The items labelled as false positives in the bottom left area have been mistakenly identified as fake positive.
* The items labelled as true Negative in the bottom right area have been accurately identified as true negative.
* True Positives are articles that have been wrongly classified as true news in the top right corner.

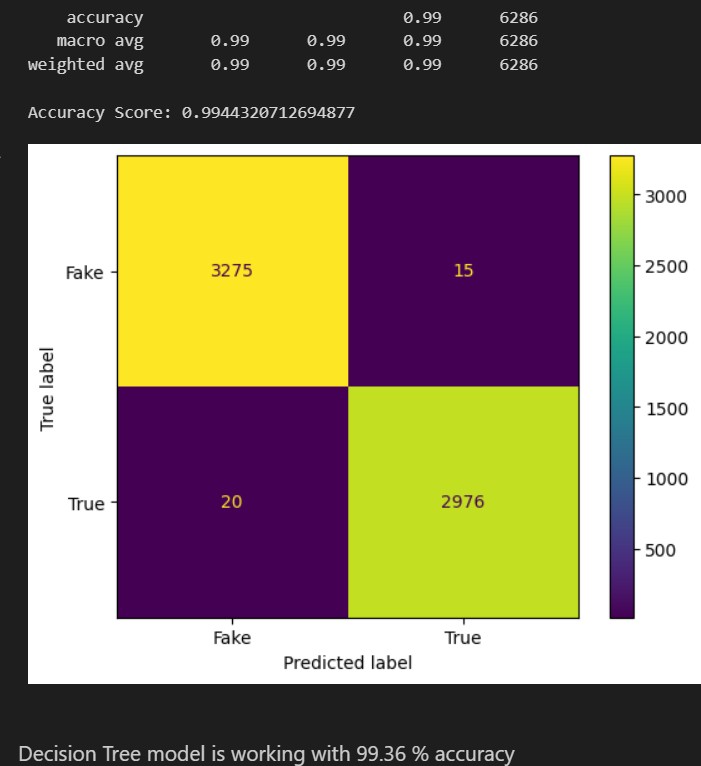
Despite the fact that XGBoost has a greater average accuracy than any other classifier employed in this dataset



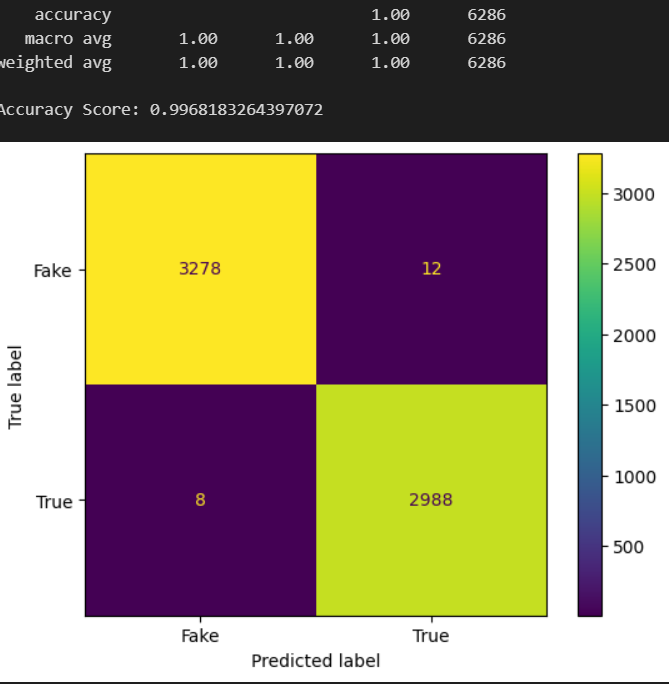
Confusion matrix of PA Classifier



Confusion matrix of Logistic Regression



Confusion matrix of Decision Tree model



Confusion matrix of XG BOOST

1. CONCLUSION

In this paper, we have proposed a unique hybrid fake news detection system that employs two types of features: linguistic and fact-verification features. The proposed detection system employs only eight features, which less compared to the stat-of-the-art approaches. It operates in two phases: training and testing. In the training phase, the detection system runs four machine learning algorithms, i.e., Logistic Regression (LR), Random Forest (RF), Additional Trees Discriminant, and XGBoost, in order to select the best

classifier for the testing phase. Evaluation results on the News data set show that the proposed detection system reaches an accuracy of 99% under XGBoost. As future work, we aim at improving the accuracy of our detection system by investigating other discriminating features such as visual- based and style-based features.Moreover, we plan to further detect other types of false information such as biased/inaccurate news and misleading/ambiguous news.

REFERENCES

1. N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi and

A. Bouras, "A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media," in IEEE Access, vol. 10, pp. 62097-62109, 2022, doi: 10.1109/ACCESS.2022.3181184.

1. Yildirim, G. A novel hybrid multi-thread metaheuristic approach for fake news detection in social media. Apple Intell (2022). <https://doi.org/10.1007/s10489-022-03972-9>
2. Thompson, R.C.; Joseph, S.; Adeliyi, T.T. A Systematic LiteratureReview and Meta-Analysis of Studies on Online Fake News Detection. Information 2022, 13,527.https://doi.org/10.3390/info13110527
3. Jarrahi, A., Safari, L. Evaluating the effectiveness of publishers’ features in fake news detection on social media. Multimed Tools Appl 82, 2913–2939 (2023). <https://doi.org/10.1007/s11042-022-12668-8>
4. Ihsan Ali, Mohamad Nizam Bin Ayub, Palaiahnakote Shivakumara, Nurul Fazmidar Binti Mohd Noor, "Fake News Detection Techniques on Social Media: A Survey", Wireless Communications and Mobile Computing, vol. 2022, Article ID 6072084, 17 pages, 2022. <https://doi.org/10.1155/2022/6072084>
5. Raturi S., Mishra A. and Maji S. (2021). Fake News Detection in Social Networks using Machine Learning: A Review. In Proceedings of the 3rd International Conference on Advanced Computing and Software Engineering - Volume 1: ICACSE, ISBN 978-989-758-544-9, pages 177-181. DOI: 10.5220/0010564800003161
6. Pandey, S., Prabhakaran, S., Reddy, N. V. S., & Acharya, D. (2022). Fake News Detection from Online media using Machine learning Classifiers. Journal of Physics: Conference Series, 2161(1), [012027]. [https://doi.org/10.1088/1742- 6596/2161/1/012027](https://doi.org/10.1088/1742-%206596/2161/1/012027)
7. Khanam, Z & Alwasel, B & Sirafi, H & Rashid, Mamoon. (2021). Fake News Detection Using Machine Learning Approaches. IOP Conference Series: Materials Science and Engineering. 1099. 012040. 10.1088/1757- 899X/1099/1/012040.
8. Sharma, Uma & Saran, Sidarth & Patil, Shankar. (2020). Fake News Detection Using Machine Learning Algorithms. 2320-2882.
9. A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A smart System for Fake News Detection Using Machine Learning," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977659.