isrjournals

International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications

Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

isrjournals

International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications

Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

Detecting Fake News on Social Media: A Hybrid Approach with Linguistic and Knowledge-Based Analysis.

Arjun. S[1], Gokulavasan MC[2], Sameer Basha SK[3], Sundhari M[4]
[123]UG Scholar–Dept CSE, GRT Institute of Engineering and Technology, Tiruttani, India.
[4*]Professor-Dept CSE, GRT Institute of Engineering and Technology, Tiruttani, India.
sankeramuniyamal@gmail. com,gokulavasan137@gmail.com, bashasameer400@gmail.com
[4*]**Corresponding Author**: **sundhari.m@grt.edu.in**

Abstract-The rapid development of different social media and content-sharing platforms has been largely exploited to spread misinformation and fake news that make people believing in harmful stories. It allows influencing public opinion, and could cause panic and chaos among population, Thus, fake news detection has become an important research topic, aiming at flagging a specific content as fake or legitimate. Fake news admired from various website are collected and that datasets are trained using Logistic regression, Random Forest Classifiers, Nave bayes, SVM and voting classifiers. Checking the dataset using XGBOOST for validation then a novel hybrid fake news detection system that combines Linguistic features and a novel set of knowledge-based features, social context base called fact-verification features. Finally real and imaginary detected using fact verification method which comprise three types of information namely, reputation of the website where the news also published, coverage opinion of well-known fact-checking websites about the news.

Keywords: Linguistic feature, novel set of knowledge-based features, fact-verification features.

1. INTRODUCTION

Fake news can significantly influence social and political landscapes, swaying elections, promoting misinformation about candidates or policies, and fostering social polarization that leads to conflict. With different groups are more susceptible to fake news, particularly young people with limited critical thinking skills and elderly individuals who may not be as proficient in navigating digital spaces. These groups are often targeted by misinformation related to health, scams, or political manipulation. Fake news tends to spread more rapidly during crises such as pandemics or natural disasters, leading to heightened panic, anxiety, and poor decision-making. For example, fake health information during the COVID-19 pandemic led to panic buying, health risks, and vaccine hesitancy. The spread of fake news undermines trust in media and public institutions, making it harder for people to discern reliable information. This misunderstanding of facts exacerbates societal issues, diverting attention from important issues and spreading confusion.

Policies define choices behavior in terms the conditions under which predefined operations or actions can be invoked rather than changing the functionality of the actual operations themselves. In today's Internet- based environments security concerns tend to increase when mobile code mechanisms are introduced to enable such adaptation, and so many researchers favor a more constrained form of rule-based policy adaptation Large-scale systems may contain millions of users and individual entities–instead, it must be possible to specify [5]. policies relating to groups of entities and also to nested groups such as sections within departments, within sites in different countries in an international organization. Policies are derived from business goals, service level agreements or trust relationships within or between enterprises.

2. METHODOLOGY

This paper proposes a hybrid fake news detection system that integrates both linguistic-based and knowledge-based analysis to effectively distinguish between legitimate and deceptive news articles. The system is structured into two primary phases: training and testing-In the training phase, the system begins with feature extraction, where a comprehensive set of features is derived from the dataset. These features include both linguistic indicators (such as word patterns, syntax, sentiment, and stylistic cues) and knowledge-based attributes (such as source credibility, cross-site publishing presence, and fact-checking site references). The extracted features are then used to train multiple machine-learning models, including Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine (SVM), and ensemble Voting Classifiers. This enables the creation of a robust fake news classification model.

During the testing phase, unseen news articles are subjected to the same feature extraction process and evaluated using the trained models. The final classification-real or fake is based on the hybrid analysis combining linguistic patterns with external fact-verification metrics. The architecture of the system ensures a multi-perspective evaluation of content, improving detection accuracy and generalizability.

3. MODULES

For this fake news detection project, we used a publicly available dataset containing social media news headlines labeled as either real or fake. The data was collected from trusted sources such as Kaggle and the LIAR dataset. The total dataset includes approximately 20,000 news headlines, with an equal number of real and fake entries.

It sounds like you've built a solid foundation for your fake news detection project by using trusted sources like Kaggle and the LIAR dataset. Having a balanced dataset with an equal number of real and fake news headlines is crucial for effective training and testing of your models. The dataset was stored in CSV format and divided into training and testing sets. Only the headline and label columns were used for model training. This approach focuses on the linguistic patterns in news

ISRJournals and Publications
Page 273

isrjournals
International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications
Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

headlines, making it suitable for social media-based fake news detection.

3.1 DATA COLLECTION

The quality of data is fundamental to the success of any machine learning model. For the fake news detection system, a balanced dataset with an equal number of real and fake news articles was collected to ensure unbiased learning. Stored in CSV format, each row represents a news article with features like the headline, article body, class label (real or fake), and potentially other metadata such as publication date or source. Before feeding the data into the model, several preprocessing steps are performed. Text is first normalized by converting it to lowercase, removing punctuation, and breaking it into tokens. Common stop words that don't contribute to meaning, such as "the" or "and," are removed. The cleaned text is then transformed into numerical form using NLP-based techniques like Count Vectorization and TF-IDF, which help capture the most informative words and phrases.

The dataset is split into a training set and a testing set. The training set helps the model learn the relationship between the text features and the class labels, while the testing set is used to evaluate the model's performance on previously unseen data, ensuring it can generalize well and is not overfitting.
To maintain balance and robustness, data augmentation methods are used when necessary. These include oversampling the minority class or modifying fake news samples by rephrasing sentences or substituting words with synonyms, which helps the model learn from a more diverse range of examples.

Fig 3.1. Data collection

3.2 DATA PREPROCESSING

To support a more effective fake news detection system, the original multi-label dataset was streamlined into a binary classification format. This process involved grouping the various original truthfulness categories into two simplified labels: True and False. The True label encompassed genuine news (such as "true", "mostly-true", and "half-true"), while the False label represented misleading or fabricated content (like "barely-true", "false", and "pants-on-fire"). This binary classification approach reduced the complexity of the task, allowing the model to focus purely on distinguishing between real and fake news. Only the news headlines were retained as inputs for the classification model. Headlines are concise and often carry the core message of a news article, making them a practical and efficient choice for this task. By excluding the full article body and additional metadata, the system was optimized for faster processing and reduced noise. approach created a lean yet robust dataset ready for training a reliable fake news detection model.

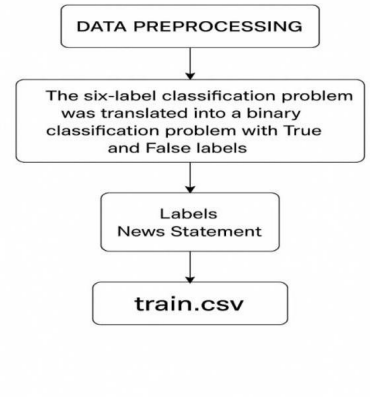

Fig 3.2 Data Preprocessing

3.3 FEATURE EXTRACTION

In natural language processing (NLP), it is essential to convert raw text into a structured numerical format that machine learning algorithms can interpret. In this project, the Count Vectorizer tool from the Scikit-learn library was utilized to transform news headlines into feature vectors. The process began by removing common English stop words—words like "is," "the," "in," and "and"—which typically do not add meaningful value to the classification task. This step helped eliminate noise and allowed the model to focus on the more informative parts of the headline. After stop word removal, the headlines were tokenized, meaning they were broken down into smaller units or tokens, using whitespace and punctuation as separators. Each token represented a word that could contribute to identifying whether a headline was real or fake.

To capture more contextual meaning, n-grams were also extracted in addition to individual tokens (unigrams). N-grams are combinations of consecutive words that often convey specific patterns or expressions, such as "tax reform" or "fake news." Including these sequences helped the model understand the relationship between words rather than just their isolated presence. Once tokenization and n-gram extraction were complete, the textual data was converted into a sparse matrix format. In this matrix, each row represented a news headline, and each column corresponded to a unique token or n-gram. The matrix values indicated the frequency with which each token appeared in a given headline. This numerical structure provided a foundation for training machine learning models, as it allowed algorithms to detect patterns in how different types of headlines were constructed. Count Vectorizer was chosen for its simplicity and its effectiveness in building a Bag-of-Words representation, which focuses on word occurrence and frequency. Although more complex techniques like TF-IDF or word embeddings

isrjournals
International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications
Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

can provide deeper semantic insight, Count Vectorizer is a strong and interpretable method, especially for tasks involving short texts like headlines. Ultimately, this feature extraction approach enabled the model to distinguish between genuine and deceptive language patterns, significantly contributing to accurate fake news detection.

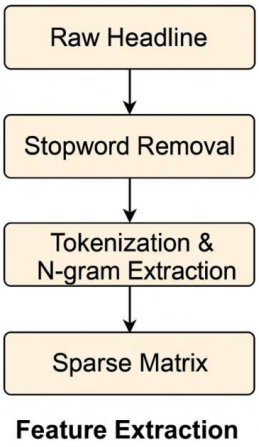

Fig:3.3. Feature Extraction

3.4 MODELING CREATION WITH XGBOOST

To build a reliable and accurate fake news detection system, a variety of supervised machine learning algorithms were applied, including Logistic Regression, Random Forest Classifier, Naïve Bayes, Support Vector Machine (SVM), and an Ensemble Voting Classifier. These models were chosen for their effectiveness in text classification and their ability to handle diverse datasets. The training process began with the use of Count Vectorizer to convert the textual data—specifically news headlines—into numerical feature vectors. These vectors captured the frequency of words and n-grams, providing a meaningful representation of the headlines that could be processed by the machine learning algorithms.

To improve the performance of each model, a comprehensive hyperparameter tuning process was carried out using Grid Search Cross-Validation. This method involved evaluating multiple combinations of parameters to identify the best settings for each model. A 5-fold cross-validation strategy was employed, where the training data was split into five parts. In each iteration, four parts were used to train the model and one part was used for validation. This process was repeated five times, ensuring that every subset of data was used for validation once, which enhanced the reliability of the results and minimized the risk of overfitting. The key objective during tuning was to maximize the F1-score, a metric that harmonizes precision and recall. This was especially important due to the potential imbalance in fake versus real news data, where accuracy alone might not reflect the true performance of the model. After the hyperparameter tuning phase, the final models were evaluated on a separate test set that was not seen during training or validation. This evaluation provided insight into how well the models would perform in real-world scenarios. Metrics such as accuracy, precision, recall, and F1-score were calculated for each model to assess their effectiveness.

In addition to the primary models, the XGBoost classifier was integrated due to its gradient boosting framework, which excels at capturing complex patterns and refining classification boundaries. Its iterative learning process contributed to higher predictive performance and further strengthened the robustness of the fake news detection system. Overall, the combination of multiple models, systematic tuning, and rigorous evaluation ensured the development of a dependable and accurate system for classifying news headlines as real or fake.

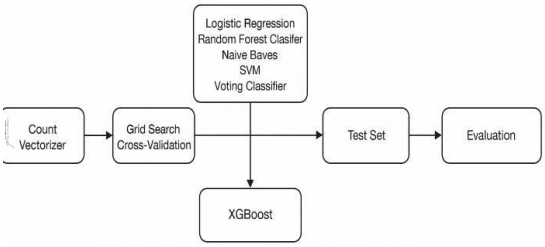

Fig 3.4. Modeling Creation

3.5. Hyperparameter Tuning

Hyperparameters are external configuration settings used to define the structure and control the learning process of machine learning models. Unlike model parameters, which are learned during training, hyperparameters are set before the training phase and play a critical role in determining model performance, generalization ability, and training efficiency. Improperly chosen hyperparameters can lead to underfitting, overfitting, or unnecessarily long training times. To optimize model performance in our study, we utilized Grid Search Cross-Validation (Grid Search CV), a systematic and exhaustive approach to hyperparameter tuning. Grid Search CV evaluates every possible combination of a predefined set of hyperparameters using k-fold cross-validation, ensuring robust and unbiased model evaluation. The best-performing set of hyperparameters is selected based on performance metrics such as accuracy and F1-score, and the final model is retrained using these optimal values on the full training set. We applied this tuning approach to all classification algorithms used in our fake news detection system.

For Logistic Regression, we varied the penalty type (l1, l2), regularization strength C, and solver (linear, saga). For SVM, we explored different values of C, kernel types (linear, rbf, poly), and gamma (scale, auto). In the case of Random Forest, we tuned parameters such as the number of estimators, maximum depth, and minimum samples required for a split. For XGBoost, we experimented with learning rate, maximum

isrjournals
International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications
Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

tree depth, number of estimators, and subsample ratios. This comprehensive hyperparameter tuning process significantly improved the predictive accuracy and robustness of our models, contributing to the overall effectiveness of our hybrid fake news detection framework.

4. ARCHITECTURE DIAGRAM

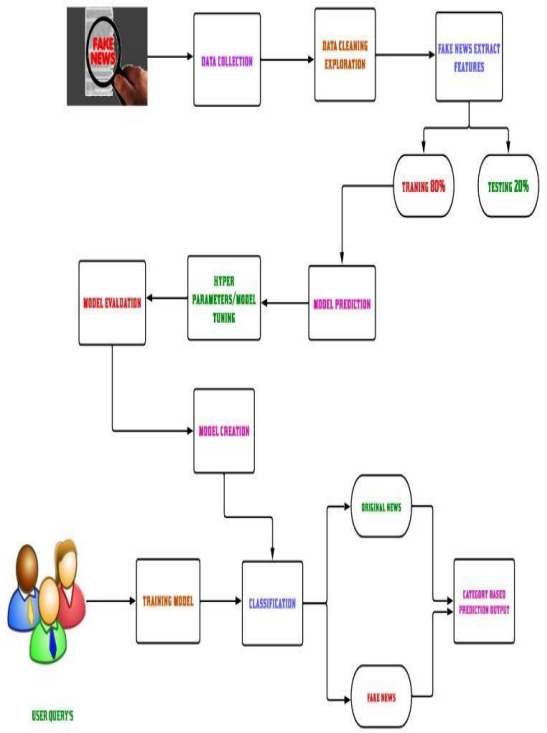

Fig 4. Architecture Diagram

5. PROPOSED SYSTEM

In this paper, we propose a hybrid fake news detection system that takes advantage of both linguistic-based and knowledge-based approaches. To the best of our knowledge, our work is the first that proposes this hybridization in the context of fake news detection. Some meta heuristics algorithms have been proposed to deal with the fake news detection issue. The proposed fake news detection system consists of two phases, namely training and testing. Both phases include a preprocessing task, which consists of cleaning and preparing the training and testing datasets of real and fake news. In the training phase, the feature extracting task extracts a set of relevant features from the training dataset, which are then fed to several machine learning algorithms to build a fake news detection model. In the testing phase, the detection model is applied on test data to decide whether the provided news articles are real or fake. Present the overall architecture of the proposed fake news detection system

.5.1 PROPOSED SYSTEM ALGORITHM

5.1.1 RANDOM FOREST ALGORITHM

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.
Please see the chart below for the evolution of tree-based algorithms over the years.

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} -f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE) (Scikit-learn only)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Fig5.1.1 Random Forest Algorithm Formula

XGBoost algorithm was developed as a research project at the University of Washington. Tianqi Chen and Carlos Guestrin presented their paper at SIGKDD Conference in 2016 and caught the Machine Learning world by fire. Since its introduction, this algorithm has not only been credited with winning numerous Kaggle competitions but also for being the driving force under the hood for several cutting-edge industry applications. As a result, there is a strong community of data scientists contributing to the XGBoost opensource projects with ~350 contributors and ~3,600 commits on GitHub.

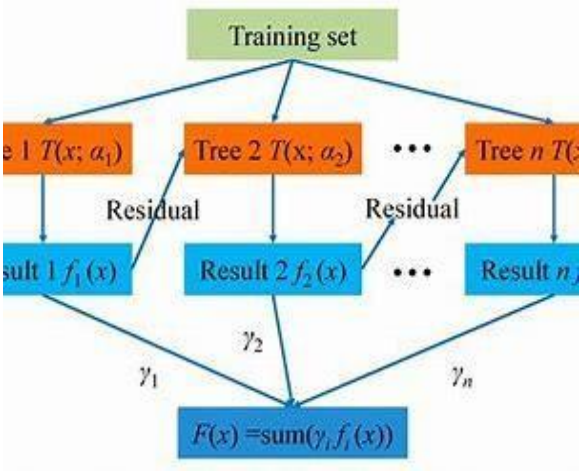

Fig5.1.2 XGBoost Algorithm Formula

The algorithm differentiates itself in the following ways:
A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
Portability: Runs smoothly on Windows, Linux, and OS X.
Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

5.2. DFD DIAGRAM

isrjournals

International Journal of Advanced Research in Computer Networking,Wireless and Mobile Communications

Volume: 5 Issue: 3 Apr,2025,ISSN_NO: 2320-7248

5.2.1 LEVEL 0:

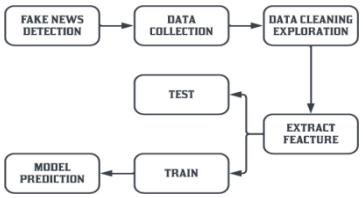

Fig5.2.1. DFD Diagram

LEVEL 1:

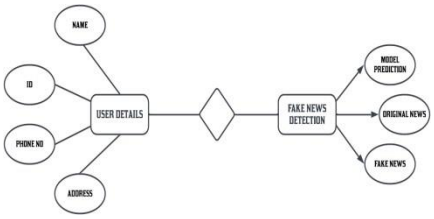

Fig 5.2.2.DFD Diagram

6.RESULT

7. CONCLUSION AND FUTURE WORK

The prosed a novel hybrid fake news detection system that employs two types of features: linguistic and fact-verification features. 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 achieves 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.

The future scope of this project is that fake news detectors can help to filter different websites that contain fake. Fake account creators are constantly adapting their tactics to evade detection. Future work could focus on developing machine learning models that can adapt to these evolving tactics and remain effective in identifying fake accounts.

8. REFERENCE

[1]. M. Park and S. Chai, "Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques," in *IEEE Access*, vol. 11, pp. 71517-71527, 2023, doi: 10.1109/ACCESS.2023.3294613.

[2]. D. Wang, W. Zhang, W. Wu and X. Guo, "Soft-Label for Multi-Domain Fake News Detection," in *IEEE Access*, vol. 11, pp. 98596-98606, 2023, doi: 10.1109/ACCESS.2023.3313602.

[3]. A. Tariq, A. Mehmood, M. Elhadef and M. U. G. Khan, "Adversarial Training for Fake News Classification," in *IEEE Access*, vol. 10, pp. 82706-82715, 2022, doi: 10.1109/ACCESS.2022.3195030.

[4]. Ravish, R. Katarya, D. Dahiya and S. Checker, "Fake News Detection System Using Featured-Based Optimized MSVM Classification," in *IEEE Access*, vol. 10, pp. 113184-113199, 2022, doi: 10.1109/ACCESS.2022.3216892.

[5]. A. Gupta *et al.*, "Combating Fake News: Stakeholder Interventions and Potential Solutions," in *IEEE Access*, vol. 10, pp. 78268-78289, 2022, doi: 10.1109/ACCESS.2022.3193670.

[6]. Y. Lin, "10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]", *My.oberlo.com*, 2021.

[7]. M. Carter, M. Tsikerdekis, and S. Zeadally, ‘‘Approaches for fake content detection: Strengths and weaknesses to adversarial attacks,’’ IEEE Internet Computer., vol. 25, no. 2, pp. 73–83, Mar. 2021.

[8]. A. A. A. Ahmed, A. Aljabouh, P. K. Donepudi, and M. S. Choi, ‘‘Detecting fake news using machine learning: A systematic literature review,’’ 2021, arXiv:2102.04458.

[9]. raşoveanu and R. Andonie, ‘‘Integrating machine learning techniques in semantic fake news detection,’’ Neural Process. Lett., vol. 53, no. 5, pp. 3055–3072, Oct. 2021