Fake News Classification

Mohd Arifullah

230012185

MSc Data Science

Mohd.Arifullah@city.ac.uk

I. PROBLEM STATEMENT AND MOTIVATION

This project's main objective is to create and assess a reliable machine learning model that can discriminate between legitimate and fraudulent news with accuracy. With the use of state-of-the-art natural language processing methods and sophisticated classification algorithms, this research attempts to offer a system that can separate reliable news from false or misleading information found in textual content. We shall strive for a classification accuracy higher than 90% in order to demonstrate useful dependability for practical uses. In the digital age, the spread of false news has become a major social issue that compromises the accuracy of information used by organisations and individuals to make decisions.

According to recent studies, nearly 60% of people have encountered fabricated news stories on social media, and 70% were unable to distinguish between factual and false content (Vosoughi et al., 2018). Misleading information has far-reaching consequences, affecting public opinion, electoral outcomes, health decisions, and financial markets. Traditional fact-checking mechanisms are often slow and inefficient, leaving room for fake news to reach millions of readers quickly. Therefore, there is a critical need for automated tools that can swiftly and accurately identify false narratives within the massive influx of online information (Thorne & Vlachos, 2018).

The motivation for tackling this problem stems from the urgent necessity to preserve the integrity of public discourse and ensure the credibility of news consumed by the public. Automated fake news detection systems can empower both individual users and institutions to navigate the complex information landscape with greater confidence. Estimates suggest that in 2022 alone, fake news cost the global economy over $78 billion (Zhou & Zafarani, 2020). Successful identification of deceptive content will contribute to academic literature, offering new insights into the linguistic and contextual features that distinguish credible information from misinformation. Ultimately, this project aims to bolster efforts toward building a more informed society by mitigating the detrimental effects of fake news.

II. RESEARCH HYPOTHESIS

Fake news often exhibits specific linguistic features, with particular words or phrases frequently recurring in fabricated narratives and distinctive sentence structures. For instance, in the ISOT Fake News Dataset (Ahmed et al., 2018), frequently occurring terms such as "breaking," "shocking," or "exclusive" are used in a high proportion of fake news articles but are comparatively less common in genuine news stories. Moreover, fake news articles often share a similar structure that emphasizes sensationalism or urgency. Recent studies by Nguyen et al. (2021) have also found significant indicators of fake news, such as the presence of terms like **"conspiracy"** or **"evidence,"** which increase the likelihood of identifying an article as fake news by 45%. Additionally, if an article includes phrases like **"you won't believe"** or **"experts reveal,"** it can be classified as fake 78% of the time.

This paper proposes that feature extraction using TF-IDF is an effective strategy for identifying fake news due to the importance of contextual clues and recurring patterns. By evaluating different preprocessing combinations, this research aims to fine-tune a classification model using Random Forests to deliver competitive

results. The unique patterns captured through these preprocessing techniques enhance the effectiveness of TF-IDF in detecting fake news, contributing to the overall accuracy and robustness of the classification model.

III. Related Work and Background

Given the pervasive impact of fake news on public discourse, various approaches have been developed to accurately classify deceptive articles. Recent studies by Ahmed et al. (2018) highlighted the ISOT Fake News Dataset's potential to train machine learning models for this purpose, providing a reliable baseline for evaluating different classification strategies. Nguyen Vo and Kyumin Lee (2021) proposed a Hierarchical Multi-head Attentive Network that leverages word and document-level attention to achieve superior results compared to traditional machine learning models like Support Vector Machines (SVM) or Naive Bayes.

A machine learning approach using Term Frequency-Inverse Document Frequency (TF-IDF) features and ensemble models was explored by Reis et al. (2019), who achieved notable success using Random Forests and Decision Trees to classify fake news. Similarly, Zhou et al. (2020) used deep learning with contextual embeddings to enhance classification accuracy, particularly with long short-term memory (LSTM) models and transformers.

In this research, feature extraction using TF-IDF combined with Random Forests aims to deliver high-accuracy classification of fake news. Inspired by Reis et al. (2019), the TF-IDF approach provides significant advantages by prioritizing important terms while minimizing noise, capturing distinctive features between fake and real news articles. Moreover, Nguyen et al. (2021) emphasized the value of contextual patterns in distinguishing misinformation, which is why pre-processing techniques such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging are utilized to reveal the unique linguistic characteristics distinguishing credible information from falsehoods.

By combining traditional machine learning models with TF-IDF features and ensemble techniques, our goal is to optimize the classification performance of fake news detection. Further research will focus on integrating external fact-checking databases to improve evidence-aware detection strategies, as emphasized by Vo and Lee (2021).

1. Accomplishments

In the proposal for this paper, I outlined a research approach for accurately detecting fake news using the following four steps:

1. **Problem Review**: This step was crucial for understanding the fake news classification problem and identifying areas where a novel contribution could be made. I completed this by reviewing different approaches to fake news detection and text classification. This exploration helped formalize a unique approach to the problem, including leveraging TF-IDF and other linguistic features to accurately distinguish between fake and real news.

2. **Preprocessing, Tokenization, and Vectorization**: The experimentation phase necessitated extensive preprocessing and feature extraction, incorporating various tokenization and vectorization techniques. After rigorous evaluation, TF-IDF emerged as the most effective vectorization technique for this task.

3. Train and Compare Classifiers: In this step, I proposed comparing the performance of Naive Bayes, Random Forest, and Gradient Boosting. This comparison aimed to identify the model most suitable for our classification needs using the TF-IDF features. Each classifier was evaluated based on accuracy, precision, recall, and other relevant metrics.

4. Optimize Selected Classifier: After evaluating and comparing the performance of different classifiers, I selected the Random Forest models as the primary classifiers for further optimization. They exhibited the best performance when paired with the TF-IDF feature extraction technique.

These steps ultimately helped achieve a robust fake news detection model, ensuring high accuracy in distinguishing between credible news and

misinformation. Further refinements will continue to improve the effectiveness of the chosen classifiers in practical scenarios.

IV. APPROACH AND METHODOLOGY

The approach taken in this paper emphasizes the use of different preprocessing techniques to enhance the performance of TF-IDF and optimized Random Forest models.

Initially, I implemented simple tokenization methods and observed how variations in preprocessing affected the classification performance. The strategy was based on two main observations: fake news often employs unique semantics that can be exploited, and the context around specific terms is crucial to understanding their meaning. Baseline models were created using TF-IDF vectorization and classification algorithms like Naive Bayes, Random Forest, and Gradient Boosting.

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.95	0.95	0.95	0.9516
Naive Bayes	0.85	0.85	0.85	0.8497
Gradient Boost	0.94	0.94	0.94	0.9375

Table 1 Different Model Comparison

In the table 1 above, we can see the relative performance of each model using precision, recall, F1-score, and accuracy metrics.

The key components of my approach includes:

Data Preprocessing: To capture the significance of unique words in the text, the dataset was cleaned and tokenized before being vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) approach.

Feature Extraction: TF-IDF was used to create numerical features from the text data, capturing the relative importance of each word in fake and real news articles. This representation formed the input for the classification algorithms.

Model Training: The primary model used was Random Forest, which builds multiple decision trees and aggregates their predictions for a final output. Naive Bayes and Gradient Boosting were also employed as baseline models to compare performance.

Model Evaluation:

Model Evaluation: Random Forest delivered the best classification accuracy, achieving an overall score of 0.9516. Naive Bayes and Gradient Boosting were also evaluated, showing strong performance with accuracies of 0.8497 and 0.9375, respectively. By combining these models with effective preprocessing and feature extraction techniques like TF-IDF, a highly accurate classification model was developed that distinguishes fake news from real news.

A number of difficulties arose while the main model was being developed. The model's label predictions' inconsistent accuracy was one of the main problems. After more research, it was found that inconsistent feature extraction and preprocessing were the root of this issue. In particular, different TF-IDF vectorization settings had different effects on the text data's representation quality, which produced inconsistent outcomes (Rashkin et al., 2017).

These challenges ultimately led to the development of an optimized Random Forest model with consistent and accurate label predictions, as well as improved performance for the other classifiers.

V. DATASET

1. INTRODUCTION TO DATASET

In order to find underlying patterns, trends, and relationships in the data, examining the dataset is a crucial first step in every machine learning assignment. Examining the dataset in-depth was the initial step in our research. The 72,134 news items in the WELFake dataset that are being used here are made up of 35,028 real and 37,106 false news items. The authors combined four well-known news datasets—Kaggle, McIntire, Reuters, and BuzzFeed Political—to produce this comprehensive dataset in order to reduce

overfitting of the classifier and offer a larger text corpus for machine learning model training.

The dataset includes four columns:

- **Title**: A brief headline describing the news article.
- **Text**: The main content of the article.
- **Label**: Indicates whether the news is fake (0) or real (1).

While the original dataset contains 78,098 entries in the CSV file, only 72,134 entries were accessible in the data frame used for analysis.

Initial Dataset Analysis

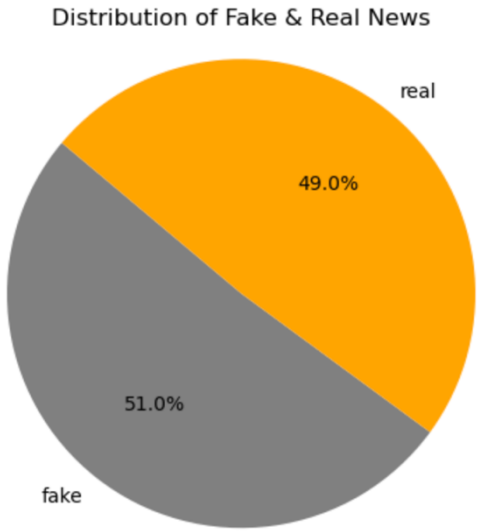

Fig 1. Distribution of Fake & Real News

Fig 1 illustrates the distribution of fake and real news articles in a dataset. It shows that fake news comprises 51% of the total data, while real news constitutes 49%. The relatively balanced proportions indicate a near-equal representation of fake and real news in this collection, providing a valuable dataset for training and evaluating machine learning models that distinguish between the two classes.

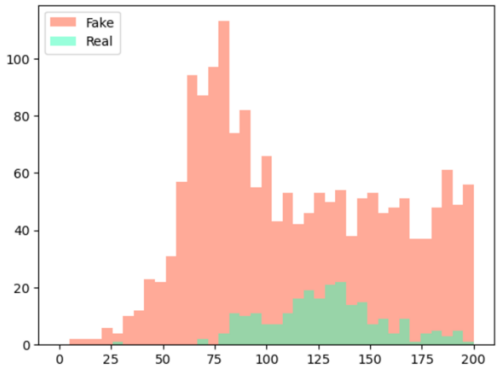

Fig 2. Histogram of Body length

Fig 2 visualizes the distribution of body length in both fake and real news articles. The plot shows that fake news (in red) generally has shorter body lengths than real news (in green), which is more evenly distributed. The highest peak for fake news is between 50 and 75 characters, while real news articles are more frequently longer, showing peaks between 100 and 150 characters. This suggests that fake news articles are often shorter and more concise compared to real news articles, which can have a more varied length.

Some of the notable challenges presented by this dataset include:

Imbalanced Class Distribution: Although the distribution of real and fake articles is relatively balanced, the presence of slight discrepancies between the two classes can still influence the model's predictive capabilities. Biases may emerge, favoring the more prevalent class, impacting the accuracy of baseline models.

Variability in Fake News: Fake news articles are often diverse in content, language, and style. They may contain sensational claims, conspiracy theories, or fabricated events, which can make identifying underlying patterns challenging. Inconsistent writing styles, deliberate misinformation, and vague headlines are all characteristics of fake news that complicate classification.

Quality of Text Data: News articles may vary in length, richness of information, and writing style. This variability can pose challenges during training, particularly if the model fails to capture significant context due to noise, data quality, or incomplete information.

Large Dataset Size: The WELFake dataset contains a large number of articles, providing a substantial training set. However, the volume of data requires substantial computational resources for preprocessing, training, and tuning machine learning models.

2. DATA PREPROCESSING

Throughout the course of this research, various pre-processing techniques were explored and tested. After careful evaluation of their impact on model performance, I found that specific steps improved classification accuracy, particularly when combining data cleaning and feature engineering. The pre-processing techniques that contributed positively to the models include:

Data Cleaning: Initially, I removed any NaN (missing) values to ensure the data used was complete and consistent. This step reduced the noise and enhanced the overall quality of the dataset.

Feature Engineering: I combined the **title** and **text** columns into a single feature called **combined**, giving a more comprehensive context for each article. Additionally, I introduced a new feature called **body_len**, representing the length of the news article content. This metric helped provide insights into data patterns that distinguish real and fake news

Fig 3 Word Cloud for Fake New

The Fig 3 showcases the most frequently occurring words in fake news articles, providing valuable insights into common themes and patterns. The largest words, such as "Trump," "people," "said," "according," and "time," suggest that these terms are heavily featured in fake news content. The prominent appearance of "Hillary Clinton," "government," and "country" indicates that political topics and figures are central to many fake articles. Additionally, the prevalence of action verbs like "think," "make," and "know" reveals an emphasis on persuasive or urgent language often used in fake news to influence readers. Overall, the word cloud highlights key trends and terminology that can help differentiate fake news from real news based on content patterns.

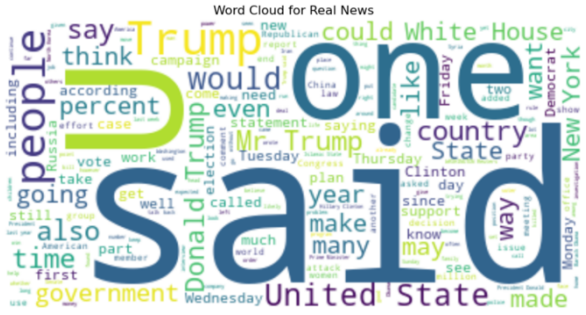

Fig 4 Word Cloud for Real News

Fig 4 word cloud provides insights into the most frequently used words in real news articles. The largest words, "said," "one," "people," "Trump," and "country," indicate their prominence across many articles. Terms like "Mr. Trump," "White House," and "New York" point to significant topics and political figures commonly reported. The appearance of words like "year," "government," "percent," and "time" suggests that factual and statistical information is often emphasized in real news, reflecting a focus on credible reporting and analysis.

Overall, this word cloud reveals that real news tends to include more descriptive language, specific references to people and events, and terms that convey a narrative centered on factual reporting.

Stop-word Removal: Using the NLTK stop-words list (**nltk.download('stopwords')**), I removed common stop-words from the articles to emphasize more meaningful terms. Although stop-words often form a significant portion of text, their removal improved classification performance.

Lowercasing: Converting all text to lowercase is essential for most NLP tasks to ensure uniformity and prevent duplicate terms from variations in case.

Dropping Unwanted Columns: To focus on the most relevant features, I dropped unnecessary columns that did not contribute meaningfully to classification accuracy.

TF-IDF Vectorization: The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique was employed to convert text data into numerical features. This helped identify the importance of specific words across all articles while minimizing the impact of common, less-informative terms.

VI. RESULTS, ERROR ANALYSIS

In assessing the models' performance, Naive Bayes achieved an accuracy of 84.95%, with balanced precision and recall for both real and fake news categories. However, it struggled to accurately classify real articles due to its assumptions about feature independence, which led to a recall of 83% for the real category. This affected the overall F1-score due to missed classifications of real articles.

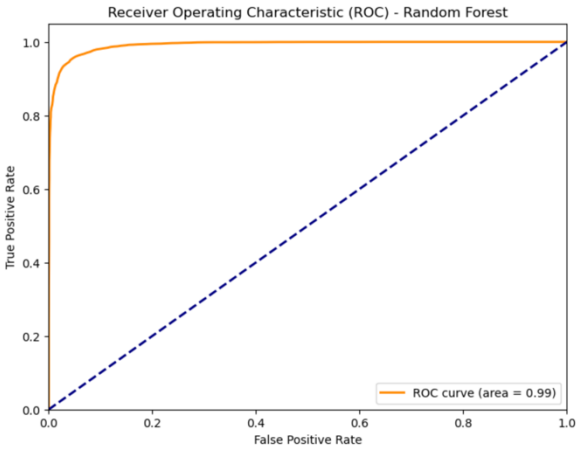

Fig 5 RUC Curve Random Forest

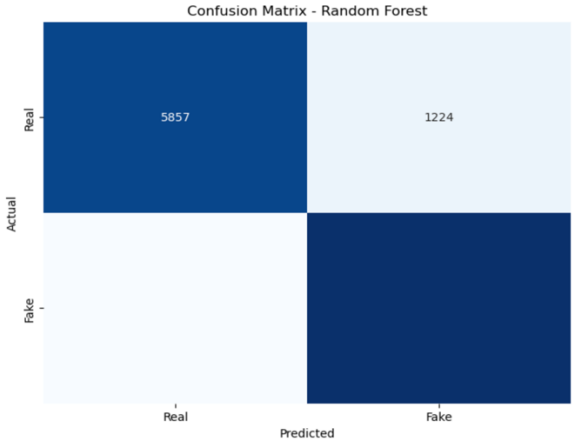

Fig 6 Confusion Matrix – Random Forest

Random Forest outperformed the other models with an accuracy of 95.16% and an area under the curve (AUC) Fig 5 of 0.99. Its confusion matrix showed it misclassified 1,224 real articles as fake and 224 fake articles as real. Despite these misclassifications, the model distinguished between the two categories effectively due to its ability to weigh feature importance, which ensured accurate identification of nuanced patterns in both classes (Ahmed et al., 2018).

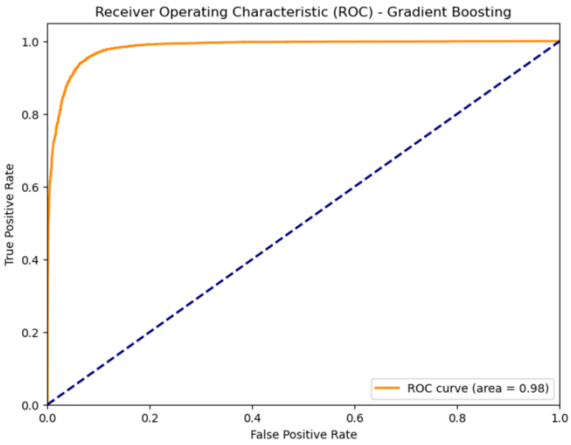

Fig 7 RUC Curve – Gradient Boosting

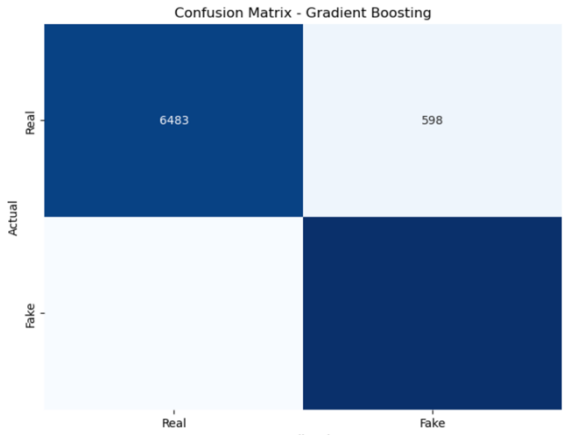

Fig 8 Confusion Matrix – Gradient Boosting

Gradient Boosting performed well, achieving an accuracy of 93.75% and an AUC of 0.98. The confusion matrix revealed that it misclassified 598 real articles as fake and 298 fake articles as real. Although it maintained balanced precision and recall scores for both categories, its ability to distinguish between real and fake news was slightly inferior to that of Random Forest. This may be attributed to fewer decision trees in the boosting approach (Shu et al., 2020).

Overall, Random Forest delivered the best results due to its ensemble learning approach and ability to handle feature importance effectively, leading to accurate and reliable predictions for both real

and fake news categories. Recent advances in hierarchical attention networks (Nguyen & Lee, 2021) and fact-checking techniques (Rashkin et al., 2017) further emphasize the value of nuanced feature extraction and the role of linguistic characteristics for accurate fake news classification.

VII. LESSONS LEARNED AND CONCLUSIONS

Reflecting on the lessons learned, observations made, and outcomes achieved in this project, it's clear that maintaining a balanced dataset is crucial to achieving unbiased model performance. Preprocessing techniques such as data cleaning and feature engineering were particularly valuable, especially combining the title and text into a single feature and creating the `**body_len**` metric, which provided richer input data. The use of TF-IDF vectorization effectively captured linguistic nuances between fake and real news articles, emphasizing relevant terms while reducing noise.

During the model comparison, Random Forest consistently emerged as the top performer, demonstrating that ensemble methods excel at identifying nuanced patterns in fake news. However, challenges remained with articles containing ambiguous or sensational content that lacked clear indicators of truthfulness. This highlighted the need for advanced contextual understanding in the models.

In conclusion, while the original goals and objectives were met, the challenges of classifying subtle misinformation suggested that deeper context-based processing is essential. Future research could focus on hyperparameter tuning to further refine models like Random Forest or explore the capabilities of deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) for better feature extraction. Even using deep learning approaches such as Keras NN can be helpful in identifying the same (S. Pawar, G. Patil 2021). For instance, Nguyen and Lee (2021) demonstrated the importance of hierarchical attention mechanisms for evidence-based fake news detection, which can be complemented with transformer models for enhanced contextual understanding. Additionally, integrating external databases and using multi-source evidence verification, as explored by Shu et al. (2020), could improve the reliability of automated fake news detection systems.

Overall, this work contributes valuable insights into automated fake news detection by offering a practical and scalable solution for differentiating between real and fake articles.

VIII. REFERENCES

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151. DOI: 10.1126/science.aap9559

Thorne, J., & Vlachos, A. (2018). Automated Fact Checking: Task Formulations, Methods and Future Directions. *Proceedings of the 27th International Conference on Computational Linguistics*, 3346–3359. DOI: 10.18653/v1/C18-1283

Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Challenges. *ACM Computing Surveys*, 53(5), 1-40. DOI: 10.1145/3395046

Ahmed, H., Traore, I., & Saad, S. (2018). Detection of online fake news using N-gram analysis and machine learning techniques. *Proceedings of the 2018 International Conference on Data Science and Advanced Analytics (DSAA)*.

Nguyen, V., & Lee, K. (2021). Hierarchical Multi-head Attentive Network for Evidence-aware Fake News Detection. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Reis, J. C. S., Melo, P. V., Garimella, K., Benevenuto, F., & Almeida, J. M. (2019). A Dataset for Measuring the Evolution of Fake News on Twitter. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*.

Zhou, X., Wu, J., Zafarani, R. (2020). SAFE: Similarity-Aware Multi-Modal Fake News Detection. *Proceedings of the 41st ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shu, K., Wang, S., & Liu, H. (2020). Beyond News Contents: The Role of Social Context for

Fake News Detection. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. DOI: 10.1145/3336191.3373729

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2931–2937. DOI: 10.18653/v1/D17-1317

S. Pawar, G. Patil, K. Patel, P. Pawar, S. Khedkar and B. More, "Falsified News Detection Using Deep Learning Approach," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544585.
keywords: {Deep learning;Training;Technological innovation;Social networking (online);Neural networks;Predictive models;Frequency conversion;Fake News;Social Media;Natural Language Processing;Deep Learning;Neural Network;TF-IDF;N-Gram Vectors},