APPLIED ARTIFICIAL INTELLIGENCE

(REPORT)

BY

PASCAL UGONNA AKANO
202445626

FAKE POLITICAL NEWS DETECTION USING SENTIMENT CLASSIFICATION

1. Introduction

The rise of political misinformation poses a significant threat to democratic institutions, public trust, and informed decision-making. With the rapid spread of unverified content on digital platforms, distinguishing between factual and fabricated political statements has become increasingly challenging. Manual fact-checking cannot keep pace with the volume of content shared daily, creating an urgent need for automated solutions.

This project addresses the NLP task of classifying political news statements as real or fake using supervised machine learning. By focusing on textual content, the study aims to contribute to the development of scalable tools for detecting misinformation and supporting digital media literacy.

2. Scope of Study

This study investigates fake political news detection using natural language processing (NLP) and supervised learning, framed as a binary classification task distinguishing real from fake statements. The dataset used is the LIAR corpus introduced by Wang (2017), comprising over 12,000 fact-checked political claims. Its concise and well-labeled nature makes it ideal for evaluating both traditional and deep learning models on short-form political misinformation.

Earlier work by Ahmed et al. (2017) demonstrated that simple lexical features like n-grams and TF-IDF, when paired with classifiers such as logistic regression and SVMs, can perform well on short news texts. More recent studies have explored advanced neural architectures. Goldani et al. (2020) applied capsule networks to fake news detection and reported performance gains over CNNs and RNNs, citing their ability to preserve semantic and spatial hierarchies. Kaliyar et al. (2021) introduced FakeBERT, leveraging contextual embeddings from BERT to outperform traditional models in similar tasks.

This project also builds on Dev and Bhatnagar’s (2024) hybrid approach, which combined Random Forest and SVM to balance generalization and decision boundary precision—an idea implemented here in the form of a hybrid RF–SVM model. The inclusion of capsule networks is similarly motivated by their dynamic routing mechanisms and suitability for sequence-based text data (Goldani et al., 2020).

By narrowing the focus to binary classification of political claims, this study avoids the ambiguity of multi-class and multimodal tasks. The aim is to ensure clarity, reproducibility, and fair comparison across diverse architectures, while contributing insights relevant to automated political fact-checking.

3. Importance of the Study

Political misinformation is a growing threat to democratic societies. Even brief exposure to fake news can erode trust in media and distort how people understand political events (Ognyanova et al., 2020). Repeated falsehoods can shape opinions, influence voting, and undermine trust in institutions.

The impact is particularly severe when misinformation targets elections, policies, or public figures. According to the Brookings Institution (2022), widespread political falsehoods can reduce civic participation and weaken confidence in democratic systems. These effects go beyond confusion, contributing to polarisation and misinformed decision-making.

Fake news often spreads more quickly than factual content, especially when it provokes strong emotions like fear or anger. Chuai and Zhao (2020) found that emotionally charged misinformation is more likely to go viral. Social media algorithms further amplify such content by prioritising engagement over accuracy.

Given the scale of this problem, there is a clear need for automated tools that can reliably detect political fake news. While many studies focus on broader misinformation, few examine short political statements, which are often harder to classify. This project addresses that gap by evaluating machine learning models on real-world political claims from the LIAR dataset

4. SMART Objectives

This study aims to develop and evaluate a machine learning approach for detecting fake political news using natural language processing. The task focuses on short political statements and compares traditional and deep learning models using a binary fake-vs-real classification scheme. The project is guided by the following SMART objectives:

- **Specific:** Build and compare machine learning models that classify short political statements as fake or real based on their text.
- **Measurable:** Target a minimum macro F1-score of 0.60 on the validation set, balancing precision and recall, and aligning with benchmarks from prior studies (Wang, 2017; Kaliyar et al., 2021).
- **Achievable:** The dataset size, available computational tools, and proven effectiveness of models like BiLSTM and BERT make this goal realistic within the academic timeframe.

- **Relevant:** The project addresses political misinformation, with implications for public trust and digital governance, and supports research into NLP-based fact-checking systems.
- **Time-bound:** The project was completed within a six-week academic period. Initial attempts with multiclass labels were refined into a binary approach after performance issues, allowing for reliable development and evaluation by the end of August.

5. Dataset

This study uses the LIAR dataset introduced by Wang (2017), a widely adopted benchmark for political fake news detection. It contains 12,836 short political statements, each manually fact-checked and labelled by PolitiFact journalists. These statements come from various sources, including debates, speeches, interviews, and social media.

Each entry includes the statement text, a truth label, speaker information, political affiliation, context, and speaker history. For this study, only the statement text and truth label were used, ensuring the models relied solely on linguistic features in line with NLP goals.

The dataset originally featured six truth labels: "pants-fire", "false", "barely-true", "half-true", "mostly-true", and "true". While this offered detailed factual gradation, models struggled with overlapping classes and imbalanced distributions. To improve performance, the task was reframed as binary classification, grouping the first three labels as fake and the rest as real. This restructuring improved label balance and model reliability.

The dataset's focus on short, real-world political statements and its detailed manual annotations make it a strong benchmark for comparing traditional and deep learning models in detecting political misinformation.

6. Exploratory Data Analysis

This section examines the structure, distribution, and quality of the LIAR dataset. The analysis helped shape preprocessing decisions and guided model design.

Class Distribution and Label Imbalance

The original six-class labels were unevenly distributed, with *half-true*,*false*, and *mostly-true* being the most frequent. *Pants-fire* had the fewest instances (Figure 1). Label encoding clarified this imbalance, prompting the use of class weights to prevent bias during training (Figure 2). Calculated weights reflected each class's relative scarcity.

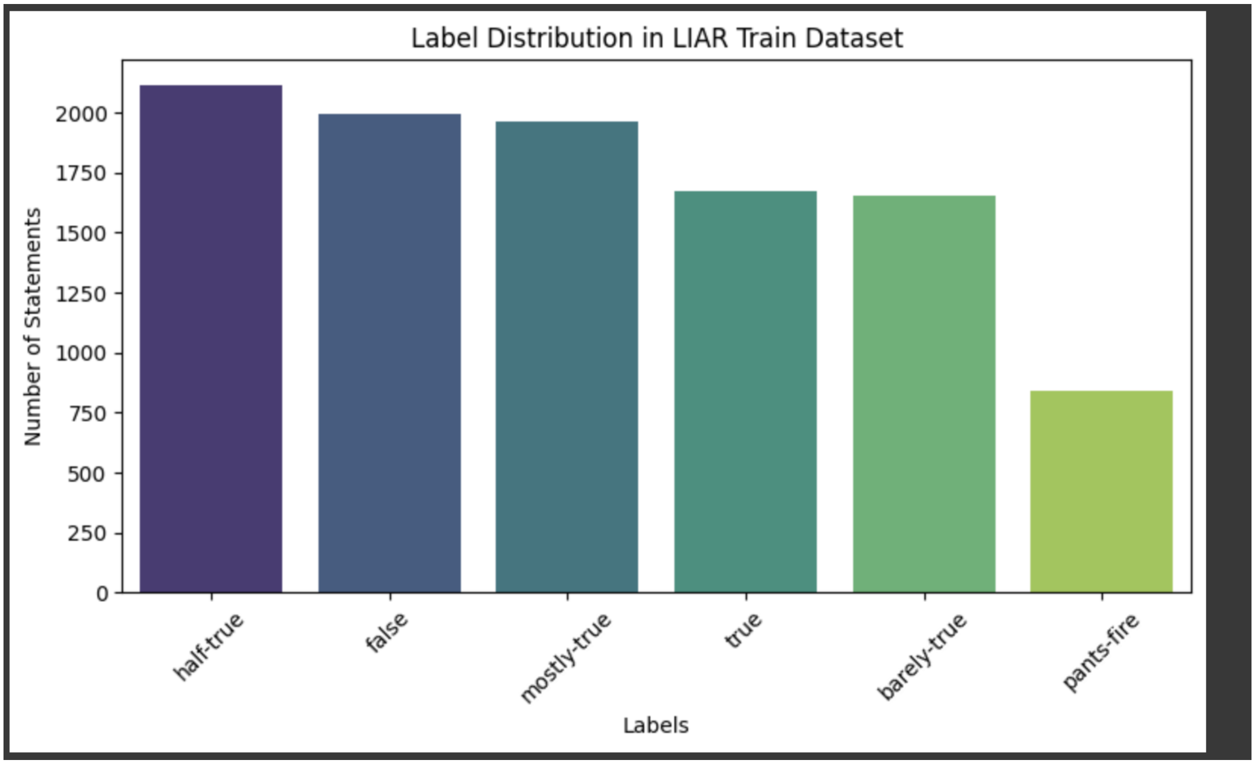

Label Distribution in LIAR Train Dataset
Number of Statements
0
250
500
750
1000
1250
1500
1750
2000
half-true
false
mostly-true
true
barely-true
pants-fire
Labels

Fig. 1

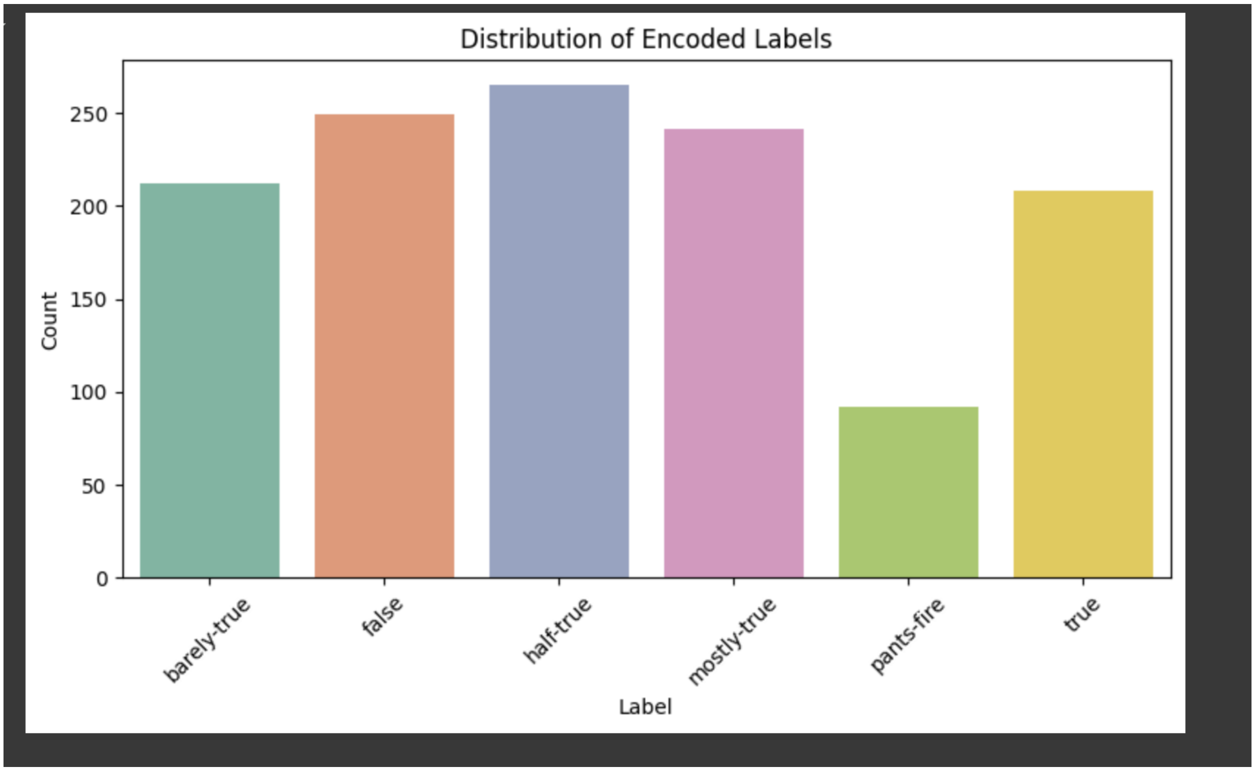

Distribution of Encoded Labels
Count
0
50
100
150
200
250
barely-true
false
half-true
mostly-true
pants-fire
true
Label

Fig. 2

Statement Length and Variability

The dataset's statements averaged 18 words, with some exceeding 400. While most labels shared similar lengths, boxplots showed that longer statements were slightly more common in *true* and *mostly-true* classes (Figures 4 and 5).

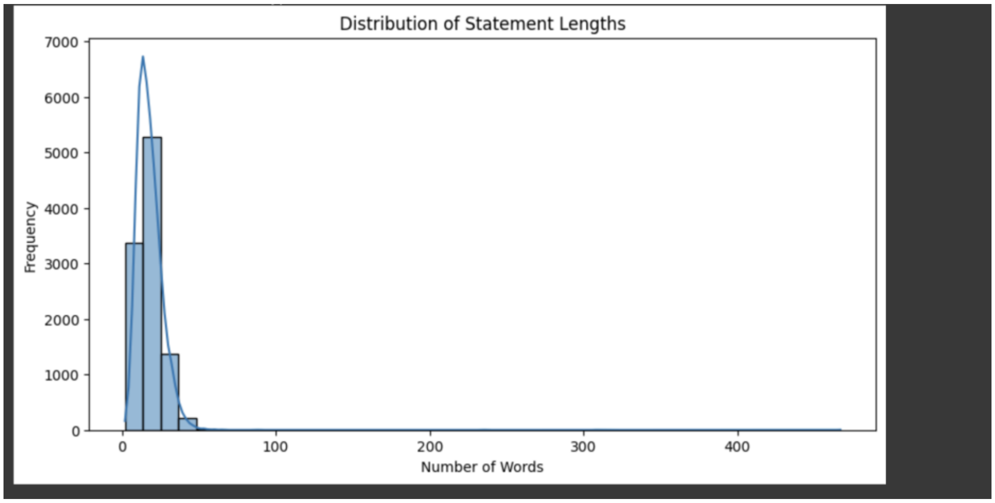


Fig. 4

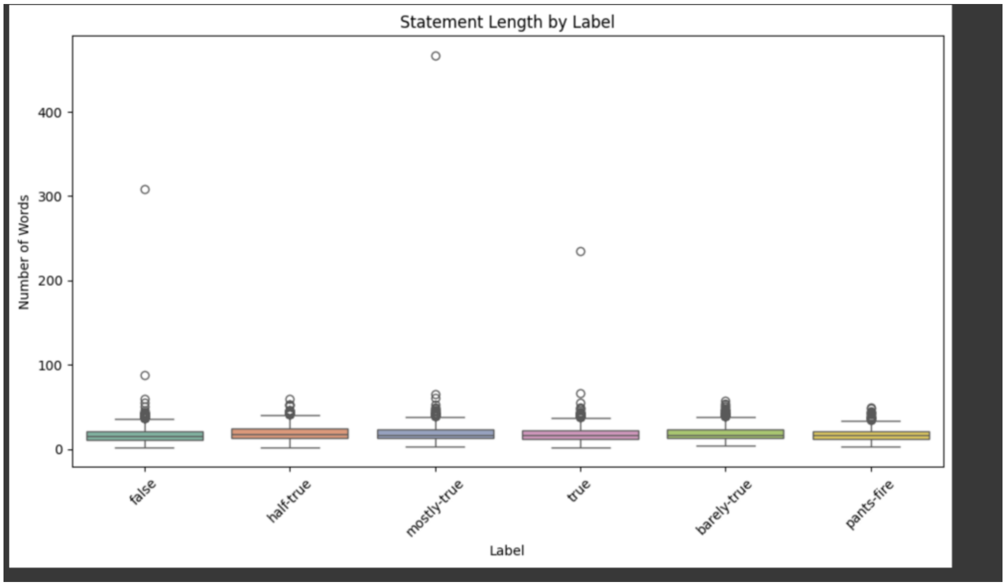


Fig. 5

Data Cleaning and Preparation

To improve input quality, missing values were addressed using logical defaults. For instance, empty *speaker* fields were replaced with "anonymous", and missing *context* values were set to "casual remark". Stopwords and punctuation were removed, and text was lowercased and lemmatised using NLTK functions. These steps ensured cleaner, more standardised inputs across the dataset.

Word Frequency Analysis

To explore semantic content, a word cloud of the cleaned training data was generated (see *Figure 6*). Prominent terms included “year,” “percent,” “people,” and “ say” reflecting the dataset’s political focus. Beyond simple frequency, text was also transformed into structured numerical form using **TF–IDF (Term Frequency–Inverse Document Frequency) vectorisation**, which quantifies the importance of words across documents. Unigrams and bigrams were retained, capped at the 5,000 most informative features, to balance expressive power and computational cost. This representation not only facilitated later model training but also offered insights into recurring themes across truthful and deceptive claims.

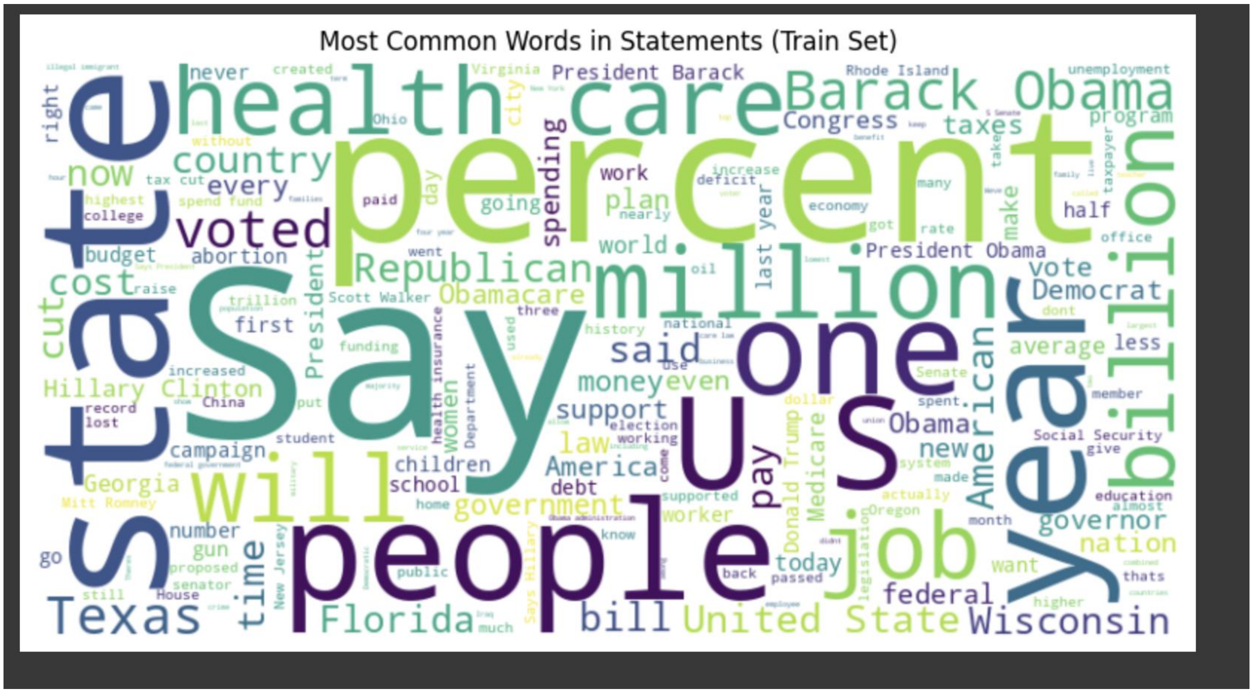
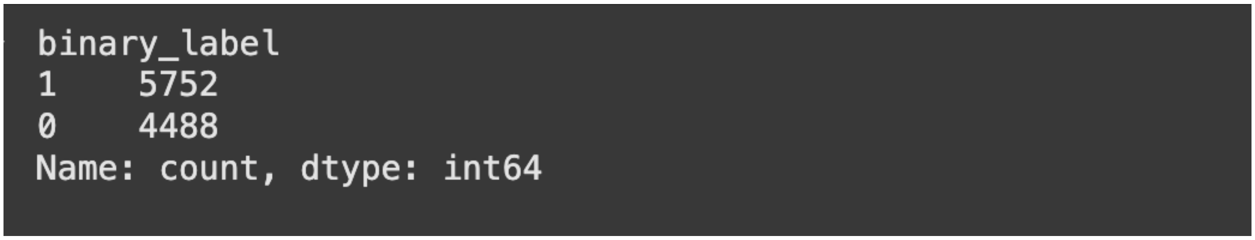


Fig. 6

Switching to Binary Classification

Given the low performance of early models in the six-class setup, labels were grouped into two categories: *Fake* (0) and *Real* (1). This improved balance, resulting in 5,752 real and 4,488 fake examples (Figure 7), and simplified the classification task.



```
binary_label
1    5752
0    4488
Name: count, dtype: int64
```

Fig. 7

7. Literature / Background review

Political fake news remains a significant challenge in the digital age. The rapid spread of misinformation has driven researchers to explore automated solutions, particularly using natural language processing (NLP). Over time, the field has progressed from traditional machine learning approaches to deep learning and transformer-based models.

Najwan et al. (2024) applied Random Forest with TF-IDF features to fake news classification and achieved 88.24% accuracy, demonstrating its strength in binary tasks. MNB remains popular for its simplicity with n-gram features, while Adeyiga et al. (2024) showed that Logistic Regression performs competitively on large news datasets. These results support the use of traditional models in this study, though they often struggle with context and subtle intent in political language.

Ahmed et al. (2017) highlighted these limitations using the ISOT dataset, noting that lexical-based models underperform on nuanced and ambiguous statements. While these methods are efficient and interpretable, they are often better suited as baselines or where computational resources are limited.

To improve performance, researchers turned to deep learning. Wang (2017) introduced the LIAR dataset, consisting of short, fact-checked political claims. Using BiLSTM models, he demonstrated that sequence-aware architectures could better capture the structure and context of political statements. Goldani et al. (2020) further explored capsule networks, which preserved spatial word relationships and improved detection of subtle misinformation. Their results showed that capsule networks outperformed CNNs and RNNs on both LIAR and ISOT datasets.

Transformer-based models such as BERT have since led the field. Pre-trained on large corpora, BERT captures deeper language patterns. Kaliyar et al. (2021) introduced FakeBERT, a fine-tuned model that achieved strong results on political fake news detection. However, transformer models are resource-intensive and require careful tuning, as seen in this project.

Hybrid models have also emerged. Dev and Bhatnagar (2024) proposed an RF-SVM combination that blends ensemble learning with margin-based classification. This hybrid approach produced robust results, and similar gains were observed in this study after hyperparameter tuning.

Interpretability remains a concern in fake news detection. Meel and Vishwakarma (2019) noted that while tools like LIME and SHAP provide insights, they often reflect only local explanations rather than the model's overall reasoning.

This study builds on these insights by testing MNB, logistic regression, Random Forest, Hybrid RF-SVM, BiLSTM, capsule networks, and BERT. The decision to shift from multiclass to binary classification also reflects strategies used in prior research to enhance clarity and model reliability.

8. Traditional Machine Learning Methods

This section presents a systematic evaluation of six traditional machine learning models applied to the task of political fake news detection. Models were tested in both **multiclass** and **binary classification** settings using **TF-IDF vectorization** to convert text into numerical features. Evaluation metrics included accuracy, precision, recall, and macro-averaged F1-score. A baseline model was first established to provide context for interpreting model performance.

Baseline Classifier

A majority-class baseline served as a performance floor. In the multiclass setup, always predicting the most frequent label resulted in 21% accuracy and a macro F1-score of just 0.07. For binary classification, defaulting to "REAL" gave 50% accuracy but failed to detect fake claims, with an F1-score of 0.28. These results highlighted the need for more robust approaches.

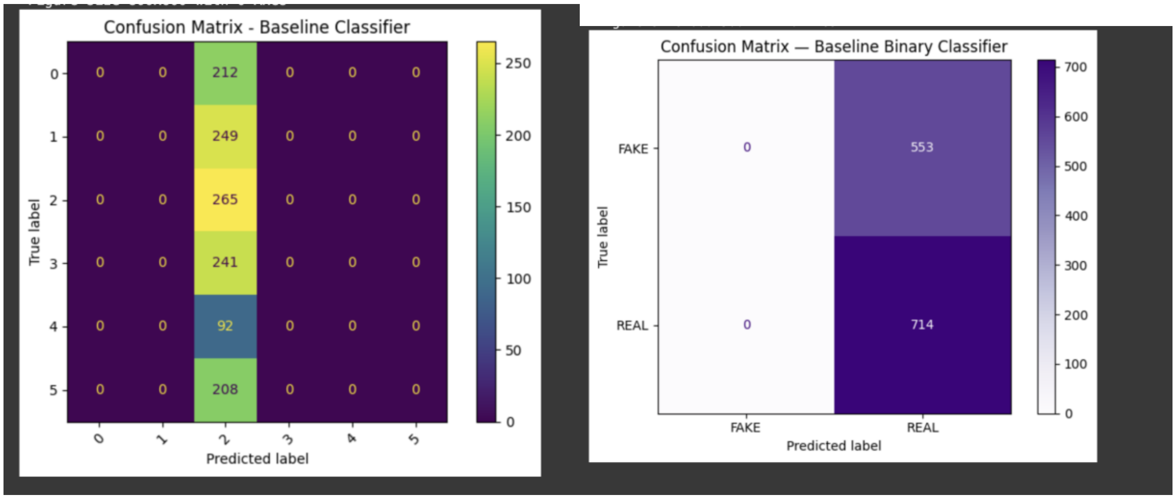


Fig. 8 Confusion matrix baseline model multiclass and binary classifier

Logistic Regression (LR)

Logistic Regression performed modestly in the multiclass setting (24% accuracy, 0.24 F1-score). Results improved in binary classification, achieving 63% accuracy and a balanced F1-score of 0.63. Hyperparameter tuning further stabilized recall and precision, particularly in detecting fake news, though overall accuracy remained consistent.

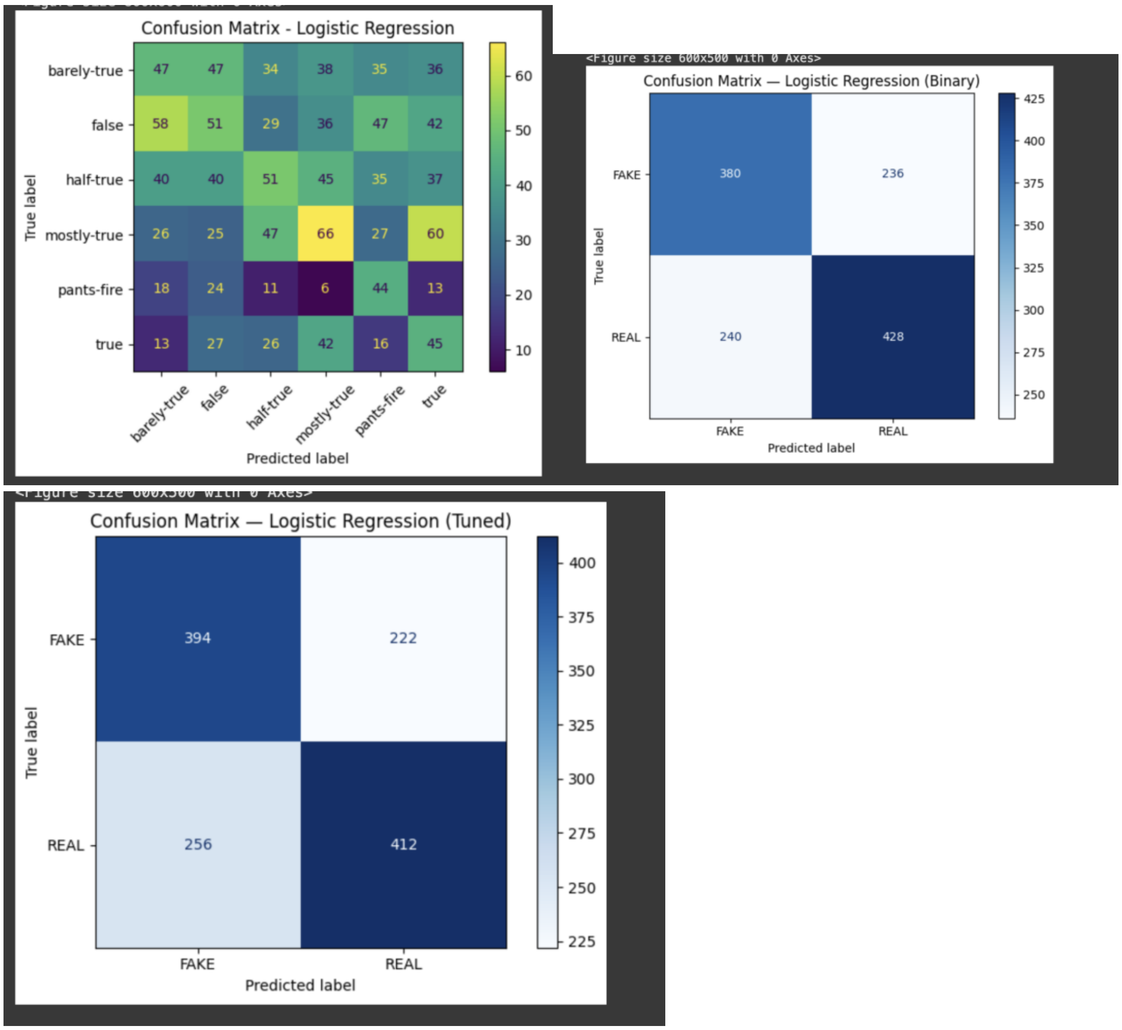


Fig. 9 Confusion Matrix logistics regression classifier

Support Vector Machine (SVM)

SVM struggled with multiclass data, yielding 23% accuracy and a 0.22 F1-score. However, in binary mode, performance rose to 62% accuracy and a 0.62 F1-score. Compared to LR, SVM offered slightly better recall for real news but was marginally weaker on precision for fake claims.

Random Forest (RF)

Random Forest outperformed earlier models in the multiclass task, reaching 27% accuracy and a 0.25 F1-score. It performed even better in the binary setting, where it achieved 64% accuracy and an F1-score of 0.63. The ensemble's ability to generalize from sparse data contributed to its balanced and reliable results.

Multinomial Naive Bayes (MNB)

MNB was the weakest multiclass performer, matching the baseline with 21% accuracy and a 0.24 F1-score. In binary classification, it achieved 61% accuracy and a 0.61 F1-score. Despite its simplifying assumptions, the model proved surprisingly effective for lightweight text classification.

Hybrid RFSVM (Random Forest + SVM)

A hybrid approach combining Random Forest and SVM was implemented to enhance performance. Random Forest-generated probabilities served as features for a linear SVM classifier. The untuned hybrid achieved 64% accuracy and a 0.63 F1-score. After tuning, results remained consistent, with slightly improved class balance. This model proved the most stable and resilient among traditional approaches.

Summary of Model Performances

Model	Multiclass Accuracy	Multiclass F1	Binary Accuracy	Binary F1
Baseline	21%	0.07	56%	0.41
Logistic Regression	24%	0.24	63%	0.63
SVM	23%	0.22	62%	0.62
Random Forest	27%	0.25	64%	0.63
Multinomial Naive Bayes	21%	0.24	61%	0.61
Hybrid RFSVM	N/A	N/A	**64%**	**0.63**
Tuned Logistic Regression	N/A	N/A	63%	0.63
Tuned Hybrid RFSVM	N/A	N/A	**63%**	**0.63**

Key Takeaways

Traditional models struggled with multiclass prediction due to label ambiguity and imbalance. However, performance improved significantly when the task was simplified to binary classification. The Hybrid RFSVM and tuned Logistic Regression models emerged as top performers, offering strong baselines for comparison with deep learning methods.

9. Deep Learning Models

This section presents our deep learning experiments for fake news detection, covering LSTM, BiLSTM, Capsule Networks, and BERT. Initial tests used the original 6-class labels, but due to poor accuracy and class overlap, we converted the task to binary classification. Grouping "pants-fire," "false," and "barely-true" as *Fake*, and the remaining labels as *Real*, resulted in a more balanced and manageable setup, consistent with prior studies.

LSTM (Multiclass)

Our initial LSTM, built with an embedding layer and a 64-unit LSTM, performed poorly on the 6-class data. Despite using dropout and early stopping, it achieved only 19.3% validation accuracy. The results underscored the challenges of multiclass fake news classification using basic RNNs without augmentation or additional features.

BiLSTM and Tuning

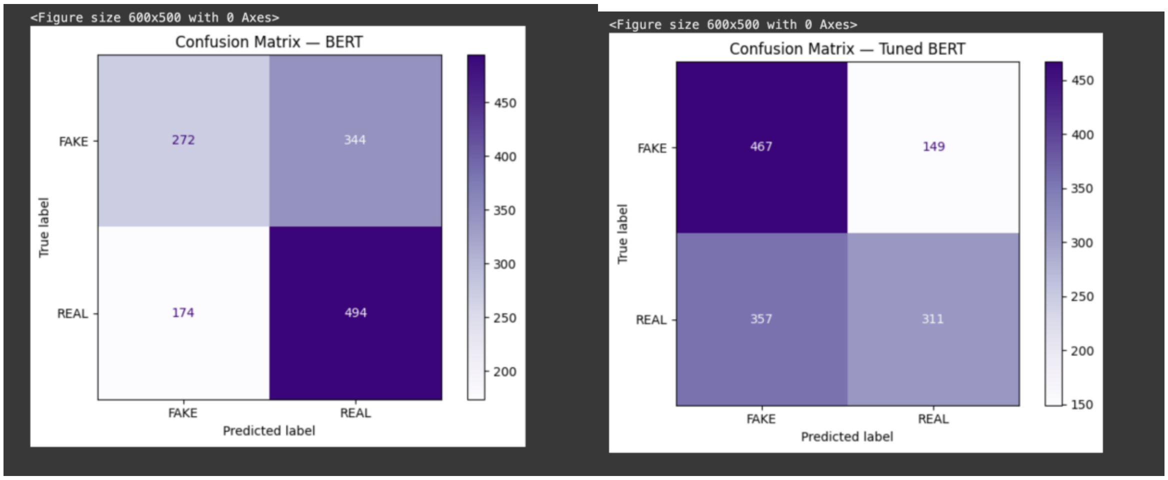
A Bidirectional LSTM was trained next, using class weights to counter imbalance. It achieved 62% accuracy and a macro F1-score of 0.61. Tuning its embedding size, LSTM units, and dropout preserved the same overall scores but led to more stable and balanced predictions across both classes.

Capsule Network

The capsule network architecture combined GloVe embeddings, convolutional layers, and capsule routing. It reached 59% accuracy and a 0.59 F1-score, showing potential but struggling with overfitting and low recall for the *Fake* class. The model's complexity may have outweighed its benefits on this dataset.

BERT (Base and Tuned)

BERT-base was fine-tuned using HuggingFace Transformers. The base model achieved 59% accuracy and a 0.59 F1-score. After tuning (learning rate: 2e-5, batch size: 16), performance improved to 60% accuracy and a 0.62 F1-score. Notably, recall for fake news reached 0.76, highlighting BERT's ability to detect misinformation.



```
<Figure size 600x500 with 0 Axes>
```

```
<Figure size 600x500 with 0 Axes>
```

Fig 10. Base BERT and Tuned BERT model

Summary

All deep models performed better under binary classification. BiLSTM and tuned BERT emerged as top performers, each with a 0.62 F1-score. BERT showed superior fake news recall, while BiLSTM offered balanced predictions with fewer resources. Capsule Networks showed promise but require deeper tuning. These results highlight the importance of task framing, contextual embeddings, and regularization in fake news classification.

10. Implementation and Refinement

This project was developed using Python in Google Colab, with libraries such as Scikit-learn, Keras, TensorFlow, and HuggingFace Transformers. The modelling process progressed from traditional machine learning to deep learning and transformer-based methods.

Preprocessing included cleaning the text, removing irrelevant characters, and applying tokenisation and lemmatisation. For traditional models, TF-IDF was used to convert text into numerical features. Initial models were trained on the original six-label dataset, but results were poor due to class overlap and imbalance. The task was therefore reframed as binary classification, grouping labels into fake and real, which improved both accuracy and training stability.

Traditional classifiers, including Logistic Regression, SVM, Random Forest, Naive Bayes, and a Hybrid RF-SVM, were first trained with default parameters. After evaluating baseline results, selected models were fine-tuned using grid search to improve performance and reduce class bias.

Deep learning models were implemented in Keras. LSTM and BiLSTM were trained with pre-trained GloVe embeddings, using padding for uniform input length. Class weighting and dropout were applied to handle imbalance and limit overfitting. BiLSTM delivered more consistent results, particularly after tuning.

A Capsule Network combining convolutional layers with capsule routing was also tested. While the model preserved spatial features, it showed signs of overfitting and would benefit from further tuning or data augmentation.

BERT was fine-tuned using the HuggingFace library. After tuning hyperparameters, the best configuration achieved a macro F1-score of 0.62, with strong performance in identifying fake news.

Overall, shifting to binary classification, balancing training data, and tuning models contributed significantly to improved results across approaches.

11. Conclusion and Discussion

Reframing the task from multiclass to binary classification had the most noticeable impact on model performance. Traditional models struggled with overlapping labels and imbalance in the original setup, while binary classification allowed for clearer separation between fake and real statements, resulting in more reliable outcomes.

Logistic Regression and Random Forest were the strongest traditional models, both achieving a macro F1-score of 0.63 after tuning. The Hybrid RFSVM offered slightly more balanced predictions. Naive Bayes underperformed due to its basic assumptions and low fake news recall, and SVM showed average results with some sensitivity to imbalance.

Among deep learning models, BiLSTM delivered balanced precision and recall, reaching an F1-score of 0.62. Tuned BERT matched this performance but stood out with a 0.76 recall for fake news, suggesting its strength in capturing nuanced language. However, BERT required more time and resources to train. Capsule Networks showed potential but were less stable and prone to overfitting.

In summary, deep learning models performed slightly better, but traditional models like Logistic Regression and Hybrid RFSVM still offered strong, interpretable results with lower computational costs. This highlights the importance of clear problem framing and thoughtful tuning over model complexity alone.

A combined approach may offer the best solution. Lightweight models could handle initial filtering, while deeper models like BERT handle more critical evaluations. Future work could include speaker metadata, larger datasets, and stress-testing models across varied political contexts.

Referencing Style Commentary

This report follows the Harvard referencing style. In-text citations use the author-date format, such as (Wang, 2017) or (Kaliyar, Goswami and Narang, 2021), with multiple sources separated by semicolons. Direct quotes, though not used in this report, would include page numbers. The reference list is ordered alphabetically by author surname and includes full publication details, including journal name, volume, page numbers, year, and DOI or stable link. Preprints and datasets are clearly identified. This style ensures proper credit and allows all sources to be easily verified for academic transparency.

References

Ahmed, H., Traore, I. and Saad, S., 2017. *Detection of online fake news using n-gram analysis and machine learning techniques*. In: *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (ISDDC)*. Vancouver, Canada, pp.127–138. https://doi.org/10.1007/978-3-319-69155-8_9

Adeyiga, J.A., Toriola, P.G., Abioye, T.E., Oluwatosin, A.E. and Arogundade, O.T., 2024. *Fake News Detection Using a Logistic Regression Model and Natural Language Processing Techniques*. *Research Square Preprint*. [online] Available at: https://doi.org/10.21203/rs.3.rs-3156168/v1

Brookings Institution, 2022. *How disinformation is undermining democracy*. [online] Brookings. Available at: https://www.brookings.edu/articles/misinformation-is-eroding-the-publics-confidence-in-democracy

Chuai, Y. and Zhao, J., 2020. *Anger makes fake news viral online*. arXiv preprint arXiv:2004.10399. Available at: https://doi.org/10.48550/arXiv.2004.10399

Dev, D.G. and Bhatnagar, V., 2024. *Hybrid RFSVM: Hybridization of SVM and Random Forest Models for Detection of Fake News*. Algorithms, 17(10), p.459. https://doi.org/10.3390/a17100459

Goldani, M., Momtazi, S. and Safaei, A., 2020. Detecting fake news with capsule neural networks. *arXiv preprint arXiv:2002.01030*. Available at: https://arxiv.org/abs/2002.01030

Kaliyar, R.K., Goswami, A. and Narang, P., 2021. *FakeBERT: Fake news detection in social media with a BERT-based deep learning approach*. *Multimedia Tools and Applications*, 80, pp.11765–11788. https://doi.org/10.1007/s11042-020-10183-2

Meel, P. and Vishwakarma, D.K., 2019. *Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities*. *Expert Systems with Applications*, 153, p.112986. https://doi.org/10.1016/j.eswa.2019.112986

Najwan, T.A., Hassan, K.F., Abdullah, M.N. and Al-hchimy, Z.S., 2024. *The Application of Random Forest to the Classification of Fake News*. BIO Web of Conferences, 97, p.00049. https://doi.org/10.1051/bioconf/20249700049

Ognyanova, K., Lazer, D., Robertson, R.E. and Wilson, C., 2020. *Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School Misinformation Review*, 1(3). Available at: https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power/

Wang, W.Y., 2017. *"Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection*. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada, pp.422–426. https://doi.org/10.18653/v1/P17-2067