The Challenge of Misinformation

Sakthi Sathya Pasupathy, Yukti Bishambu

Introduction and Motivation

The proliferation of social media and online platforms has transformed how information is shared, allowing news to spread rapidly across large audiences. However, this convenience comes with a significant downside: the accelerated spread of misinformation. Fake news has the potential to mislead the public, disrupt social harmony, influence political outcomes, and create widespread panic. As a result, combating fake news has become a critical challenge in today's digital world. Traditional fact-checking methods, while important, cannot keep up with the speed and volume of online content. Consequently, there is a strong motivation to explore automated techniques that can detect fake news accurately and efficiently. This project is driven by the goal of applying machine learning and deep learning methods to build a reliable fake news detection system. Such a system could assist fact-checkers, support media platforms in maintaining information integrity, and ultimately help restore public trust in the media.

Problem Statement

The core problem addressed in this project is the automatic classification of news statements into categories based on their truthfulness. Specifically, the project focuses on identifying whether a given statement is true, half-true, or false, using both the textual content and relevant metadata such as speaker information and political affiliation. Two key research questions guide the project. Firstly, which machine learning approaches are most effective in detecting fake news based on linguistic patterns? And secondly, apart from the text itself, what additional features can improve the accuracy of fake news detection? The project seeks to develop models that not only rely on sophisticated language understanding but also leverage structured data to enhance prediction performance.

Literature Review

The LIAR dataset, introduced by Wang (2017), comprises 12.8K manually labeled short political statements from PolitiFact, each annotated with six fine-grained truth labels based on their authenticity. Each statement is accompanied by metadata features such as the topic, context, speaker, speaker's job, state, party affiliation, and speaker's credit history. This dataset has become a benchmark for evaluating fake news detection models.

Wang's initial work proposed a hybrid Convolutional Neural Network (CNN) model that integrated textual and metadata features, demonstrating improved performance over text-only models. This model achieved an accuracy of 27.4%, outperforming traditional models such as SVM and Logistic Regression, which hovered around 25%. Although the accuracy might seem low in absolute terms, it was a significant improvement over chance performance (approximately 16.7% for six classes). It reflected the complexity of the six-class classification problem and set a benchmark for future models to improve upon.

Building upon Wang's work, later studies explored different model architectures and feature representations. For example, Goldani et al. (2020) applied capsule neural networks to the LIAR dataset, achieving a modest improvement over Wang's CNN by capturing spatial hierarchies within the data. Whitehouse et al. (2022) went a step further by incorporating external factual knowledge from Wikidata into pre-trained transformer models, which helped boost performance on

fact-checking tasks. Their knowledge-enhanced BERT model significantly improved fake news detection on the LIAR dataset, emphasizing the value of incorporating external factual information.

Expanding on this, Upadhayay and Behzadan (2020) introduced Sentimental LIAR, an enriched version of the original dataset that includes sentiment and emotion features for each statement. Their fine-tuned BERT-based model achieved over 70% accuracy on a three-class version of the LIAR dataset, demonstrating that emotional tone and sentiment can provide additional cues for detecting deception and exaggeration in political statements.

Yang et al. (2022) introduced CofCED, a Coarse-to-fine Cascaded Evidence-Distillation neural network designed for explainable fake news detection. By leveraging raw reports instead of relying solely on fact-checked data, CofCED employs a hierarchical encoder and cascaded selectors to identify and distill the most pertinent evidence, enhancing both detection performance and the quality of explanations. This model exemplifies the shift towards more transparent and effective fake news detection methodologies.

Early CNN-based models established the foundation for fake news detection, but recent advances using transformers, metadata, and external knowledge have significantly improved performance. Despite this progress, challenges remain, highlighting the need for continued innovation in model architecture and feature engineering.

Methodology

Data Pre-processing

After initial exploration, the six classes were simplified into three broader categories: True, Half-True, and False. The label distribution shows that 44.2% of the statements are classified as False, making it the most common class, followed by 35.3% True and 20.6% Half-True statements. This indicates a moderate class imbalance, with a higher prevalence of false information in the dataset.

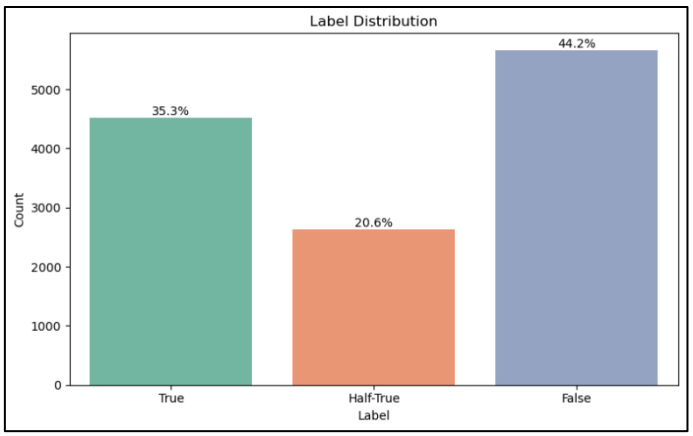

Figure 1: Simplified Label Distribution

The exploratory data analysis revealed key insights into the distribution of statements across different topics and speakers. Healthcare emerged as the most discussed topic, and notably, over half of the statements related to healthcare were labeled as false, highlighting it as a major area of misinformation.

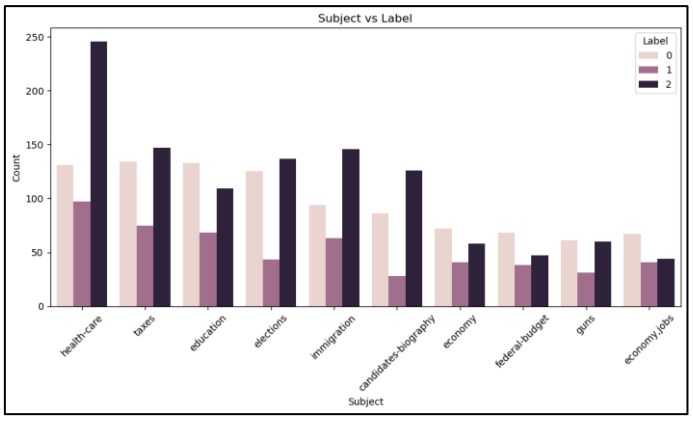

Figure 2: Label Distribution across subjects

When examining the speakers, Barack Obama accounted for the maximum number of statements, with almost 50% of his statements being labeled as true. These observations suggest that certain subjects like healthcare are more prone to misinformation, and the credibility of statements can vary significantly depending on the speaker. Such patterns are crucial for guiding feature selection and model training in fake news detection.

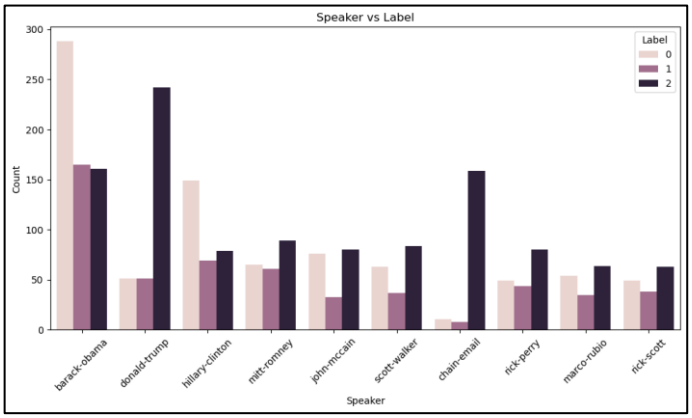

Figure 3: Label Distribution across speakers

Data preprocessing was an essential first step. Missing values were found in several columns, and complete records with nulls across all features were removed. For features with partial missing values, such as "Job", "State", and "Context," the missing entries were imputed with the label "Missing" to retain as much data as possible. Text preprocessing involved converting all statements to lowercase, removing punctuation and stopwords (commonly used words that add little meaning), and applying tokenization and stemming to standardize the text input.

Feature engineering was crucial for enhancing model performance. Term Frequency–Inverse Document Frequency (TF-IDF) was applied to the cleaned statements, converting the textual information into a numerical form, enabling classical machine learning models to process linguistic patterns effectively.

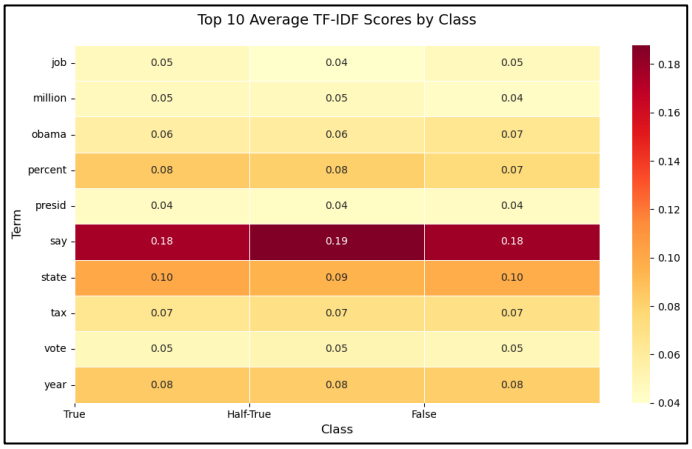

Figure 4: Top 10 terms by average TF-IDF score across classes

TF-IDF measures the importance of a word relative to a document and the entire corpus, helping to highlight discriminative words. In addition, new features were engineered, such as the length of the statement, the total number of statements made by the speaker, and the distribution of the speaker's past statements across the truthfulness categories.

Model building

Several models were then trained and evaluated. Logistic Regression and Support Vector Machine (SVM) were used as baseline models. Both models utilized TF-IDF features combined with structured metadata, offering fast training times and interpretability. Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN), was implemented to capture sequential dependencies within the text. Bidirectional Encoder Representations from Transformers (BERT) was employed to exploit contextual understanding of the statements without manual feature engineering. These models relied solely on raw text due to computational constraints.

Evaluation was conducted using multiple classification metrics such as overall accuracy, precision (correctness of positive predictions), recall (ability to capture all relevant instances), F1-score (harmonic mean of precision and recall), and confusion matrices, with a special focus on performance metrics specific to the 'false' class, which is crucial for real-world fake news detection scenarios.

Results

The results of the project (Table 1) revealed that traditional machine learning models performed competitively compared to deep learning models for fake news detection. Logistic Regression achieved an accuracy of 63%, while the Support Vector Machine (SVM) achieved a comparable accuracy of 62%. Both models demonstrated a strong F1-score of approximately 70% specifically for the "False" class, which was a critical focus of this study. This highlights that, with appropriate feature engineering, simpler linear models can be highly effective for tasks involving short text statements. Their ability to

generalize well despite the relatively straightforward structure of the input data also made them attractive choices, especially in resource-constrained environments where computation speed and interpretability are important factors.

Model	Accuracy	Recall (F)	Precision (F)	F1-Score (F)
Logistic Regression	63%	75%	66%	**70%**
Support Vector Machine	62%	73%	66%	**70%**
LSTM (10 epochs)	55%	67%	61%	64%
BERT (1 epoch)	51%	**80%**	50%	62%

Table 1: Results Summary

On the other hand, deep learning models showed mixed performance. The Long Short-Term Memory (LSTM) network, even after being trained for 10 epochs, only achieved an accuracy of 55% and an F1-score of 64% for the "False" class. Although LSTM architectures are designed to capture long-term dependencies in sequential data, their effectiveness was limited in this case. This can be attributed to the nature of political statements in the dataset, which tend to be short and contextually sparse and may not benefit significantly from sequential modeling. Moreover, deep learning models are generally data-hungry, and the relatively small dataset size of around 12K records may have hindered their ability to generalize well.

Interestingly, the BERT model showed a different trend. Even after training for just one epoch, BERT achieved a recall of 80% for the "False" class, indicating its strength in identifying deceptive statements. However, BERT's overall precision was lower, suggesting that while it successfully flagged many false statements, it also misclassified a significant number of true statements as false. This trade-off between recall and precision is crucial depending on the application: in high-stakes misinformation detection, maximizing recall (catching as many falsehoods as possible) is often prioritized over precision.

Conclusion

The project demonstrated that both classical machine learning models and modern deep learning architectures possess distinct strengths in the task of fake news detection. Logistic Regression and Support Vector Machines (SVM) proved to be effective, interpretable, and resource-efficient, particularly when enhanced with structured metadata features. In contrast, BERT highlighted the potential of transformer-based models to capture nuanced linguistic patterns and achieve higher recall rates, although at the expense of increased computational complexity and training time.

While deep learning models show significant promise for advancing misinformation detection, classical models continue to offer strong and reliable baselines, especially in scenarios where computational resources are limited. Ultimately, the choice between classical and deep learning approaches should be guided by the specific goals of the application, whether prioritizing interpretability, computational efficiency, or maximum detection accuracy.

Future directions include fine-tuning BERT more extensively, incorporating data augmentation techniques to improve model robustness, and developing cascading models that can analyze not just isolated statements but the spread of misinformation across networks. Hybrid architectures such as combining BERT embeddings with structured metadata features could offer a promising path forward by bridging the gap between accuracy and contextual nuance. By integrating these advancements into fact-checking platforms, automated tools could significantly curb the spread of false information while maintaining scalability.

This project represents an important step toward addressing the much larger and more complex problem of

misinformation in the digital age. While the models developed here can aid in detecting false content more efficiently, they are only part of a broader solution. Combating misinformation at scale will require fostering a culture of critical thinking, responsible information sharing, and proactive platform governance. This effort marks an early but meaningful step toward combating the misinformation epidemic and creating a digital environment where truth is amplified and misinformation is swiftly recognized and contained.

References

1. Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, 422–426. https://aclanthology.org/P17-2067
2. Goldani, M., Safabakhsh, R., & Momtazi, S. (2020). Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101, 106991. https://doi.org/10.1016/j.asoc.2020.106991
3. Whitehouse, C., Weyde, T., Madhyastha, P., & Komninos, N. (2022). Evaluation of Fake News Detection with Knowledge-Enhanced Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1), 1425–1429. https://doi.org/10.1609/icwsm.v16i1.19400
4. Upadhayay, S., & Behzadan, V. (2020). Sentimental LIAR: Extended corpus and deep-learning models for fake claim classification. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*. *10.1109/ISI49825.2020.9280528*
5. Yang, Z., Ma, J., Chen, H., Lin, H., Luo, Z., & Chang, Y. (2022). A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection. *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)* (pp. 2608–2621). https://aclanthology.org/2022.coling-1.230/

Credits

This project was completed as part of the course MGT 6314: Understanding Markets with Data Science at Georgia Tech. The work was a collaborative effort by Sakthi Sathya Pasupathy and Yukti Bishambu. We would like to thank Professor Chris Gu for their guidance throughout the course. We also acknowledge the open-access datasets and prior research contributions which informed our literature review and model development. We used a variety of Python libraries for data preprocessing, modeling, and evaluation, including pandas, NumPy, scikit-learn, NLTK, Matplotlib, Seaborn, and the Hugging Face Transformers library with PyTorch for fine-tuning BERT-based models. This project also benefited from course lecture videos, feedback from the professor, and clarification and debugging support from publicly available AI tools such as ChatGPT, Perplexity, etc.

Appendix

The confusion matrices further illustrated that classical model like Logistic Regression and SVM misclassified fewer true statements as false compared to LSTM, which struggled with long-range dependencies in text.