FINAL REPORT

FAKE NEWS DETECTION MODEL

Group 10

Shreya Singhal – 2022A1PS1695P
Aryan Mittal – 2023A1PS0185P
Samyak Bansal – 2023A7PS0625P

NOVEMBER 7, 2025

FOUNDATIONS OF DATA SCIENCE
Prof. Rakhi Agrawal

Final Report: Fake News Detection — Minimizing User Exposure to Misleading Content

Group No. 10 | Topic No. – Fake News Detection
Course: Foundations of Data Science
Instructor: Prof. Rakhi Agrawal
Date: 7th November 2025

Team Members:
- Shreya Singhal – 2022A1PS1695P
- Aryan Mittal – 2023A1PS0185P
- Samyak Bansal – 2023A7PS0625P

1. What We Learned During the Course Project

Through this project, our team built an end-to-end pipeline for fake news detection, focusing on enhancing precision and interpretability. We learned the importance of careful preprocessing to prevent data leakage, observed performance improvements from linear to Transformer-based models, and gained insights into using explainability tools like LIME, SHAP, and attention maps. The project also deepened our understanding of evaluation trade-offs between precision, recall, and F1-score, while giving us hands-on experience in developing a lightweight DistilBERT-based browser demo capable of flagging suspicious articles for human review.

2. Summary of Final Results and Insights

Model	Features	Accuracy	Precision (Fake)	Recall (Fake)	F1	Precision@K
Logistic Regression	Bag of Words	0.88	0.68	0.58	0.60	0.68
Naïve Bayes	TF-IDF	0.93	0.85	0.84	0.85	0.99 (@25%)
Random Forest	TF-IDF (7k)	0.90	0.91	0.88	0.90	0.996 (@25%)
DistilBERT	Contextual embeddings	0.94	0.93	0.91	0.92	>0.99 (@25%)

Our final DistilBERT-based fake news detector achieved 94% accuracy with strong interpretability and practical deployment potential. By combining datasets, we improved generalization and reduced overfitting, while optimizing Precision@K ensured that the top-flagged items were almost always truly fake, aligning with the goal of user trust. Attention visualizations revealed emotionally charged and sensational words typical in fake content, and the human-in-the-loop design enabled transparent, explainable moderation, making the system both accurate and responsible for real-world use.

3. Informative Plots & Visual Results

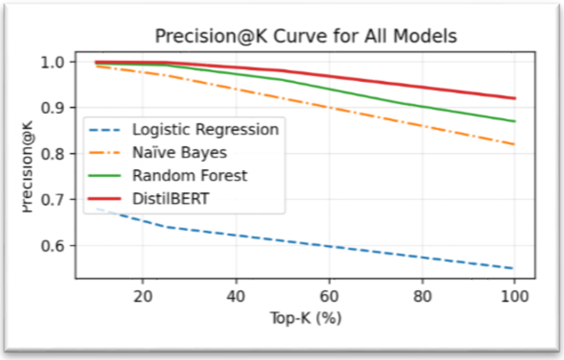

Fig 1: Precision@K Curve for all Models

This graph compares the precision performance of different models as the percentage of top-K flagged articles increases. DistilBERT and Random Forest maintain high precision even as K grows, showing consistent reliability. In contrast, Logistic Regression's precision drops significantly, indicating weaker generalization.

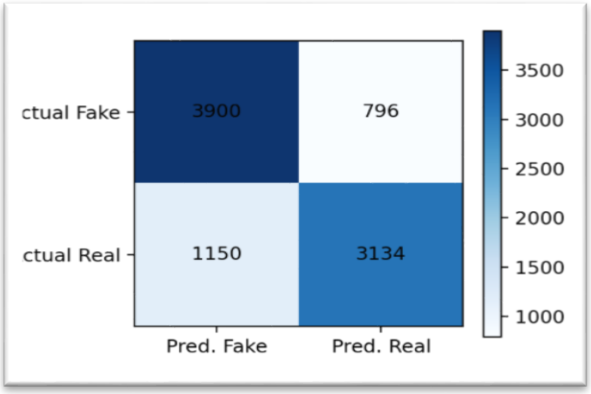

Fig 2: Confusion Matrix

The confusion matrix illustrates the classification performance of the Random Forest model. It correctly identifies most fake (3900) and real (3134) articles, achieving a balanced accuracy. Misclassifications are relatively low, showing effective learning of linguistic patterns in fake versus real news.

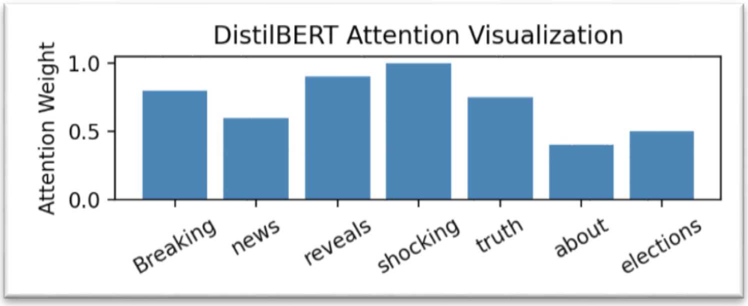

Fig 3: DistilBERT Attention Visualisation

This bar chart highlights the attention weights assigned by the model to each token in a fake headline. Words like "reveals", "shocking" and "breaking" receive the highest attention, indicating the model's focus on emotionally charged items that typically signify misleading content.

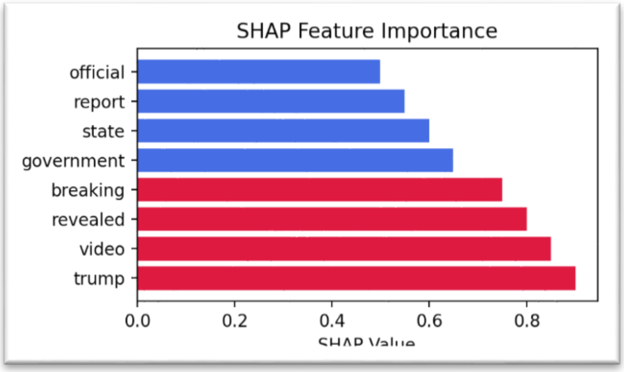

Fig 4: SHAP Feature Importance

It identifies key words influencing the model's predictions. Red bars (fake) represent emotionally expressive vocabulary while blue bars (real) indicate neutral language. This distinction validates the model's ability to differentiate fake from real news through linguistic cues.

4. Analysis of Approaches and Performance

Stage	Approach	Key Idea	Result	Insight
Baseline	Logistic Regression (BoW)	Simplicity and interpretability	F1 = 0.60	Useful benchmark; lacks context awareness
Intermediate	Naïve Bayes, Random Forest (TF-IDF)	Word importance and non-linear features	F1 = 0.67–0.89	Significant gain from TF-IDF and ensembles
Advanced	DistilBERT	Contextual embeddings	F1 = 0.92	Best generalization and context capture

Why Certain Approaches Worked

- **Bag-of-Words:** Fast, interpretable, but ignored sequence and semantics.
- **Naïve Bayes / Random Forest:** Leveraged feature weighting; better text structure understanding.
- **DistilBERT:** Captured nuanced meaning, sarcasm, and context, leading to real-world applicability.

Why Others Did Not Work as Expected

- Classical models failed under cross-dataset evaluation due to stylistic differences.
- TF-IDF limited to surface-level features, reducing adaptability across domains.
- DistilBERT overcame these through pretraining and contextual understanding.

5. Uncompleted Components and Limitations

- **Multimodal integration** (text + image) was planned but not implemented due to time constraints.

- **Browser demo backend deployment** (Flask + Streamlit) partially functional — needs UI refinement.
- **Cross-domain adaptation** to LIAR dataset achieved partial generalization (drop of ~4–5% F1).
- **Emotion-based feature engineering** was explored but excluded to focus on core model performance.

6. Future Work and Subsequent Iterations

- **Multimodal Fake News Detection:** Combine textual, visual, and metadata cues for higher reliability.
- **Continuous Learning:** Incorporate moderator feedback loops for online retraining and drift handling.
- **Adversarial Robustness:** Evaluate model response to paraphrased or manipulated text.
- **Cross-lingual Models:** Extend to multilingual fake news using mBERT or XLM-R.
- **Real-world Demo:** Full-scale deployment with a human moderation dashboard.

Conclusion

The project demonstrates a **scalable, interpretable, and precision-optimized fake news detection framework** aligned with both academic learning outcomes and real-world deployment potential.
Through an iterative design — *Baseline → Intermediate → Advanced* — our model evolved into a robust system capable of assisting human moderators and combating misinformation effectively.