Combatting Misinformation: Semantic Classification for Fake News Detection

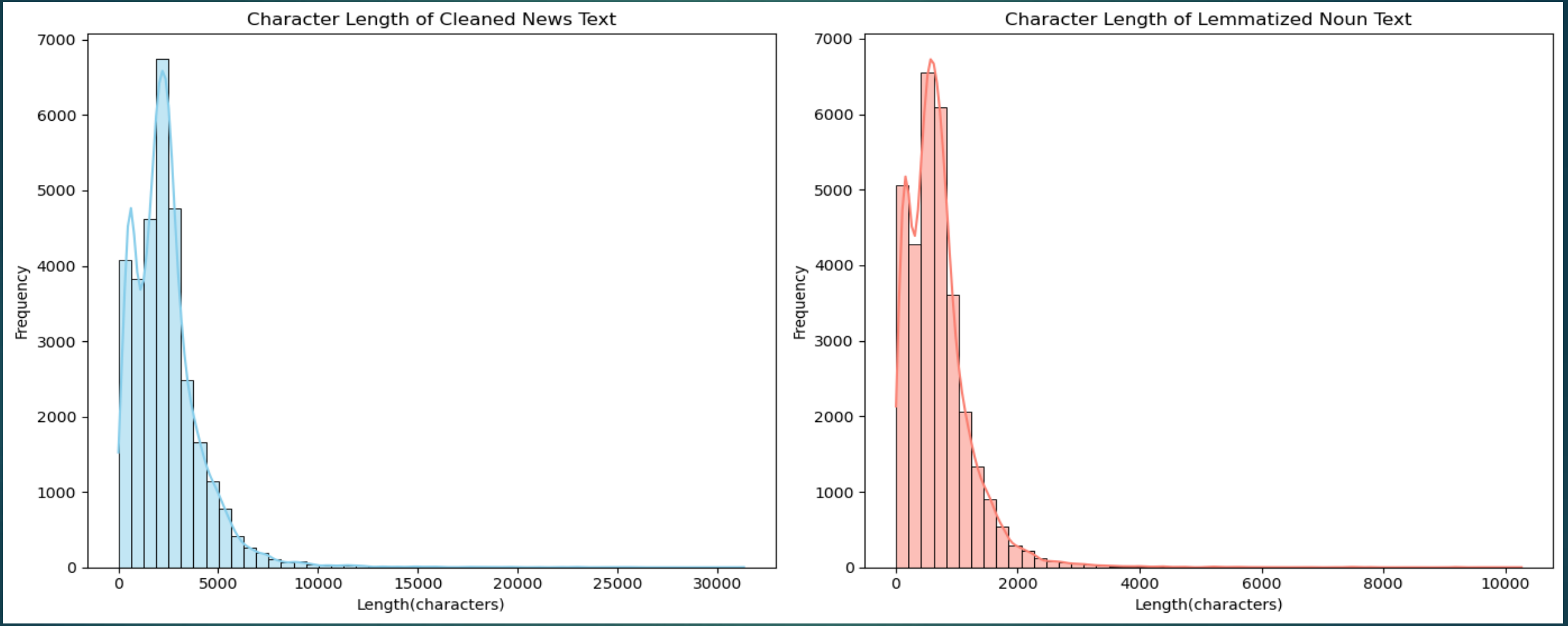
By:
Sakshi Gupta

The Challenge of Digital Misinformation

- **The Business Problem:** The viral spread of fake news damages brand reputation and social trust, creating a need for automated, real-time verification systems.

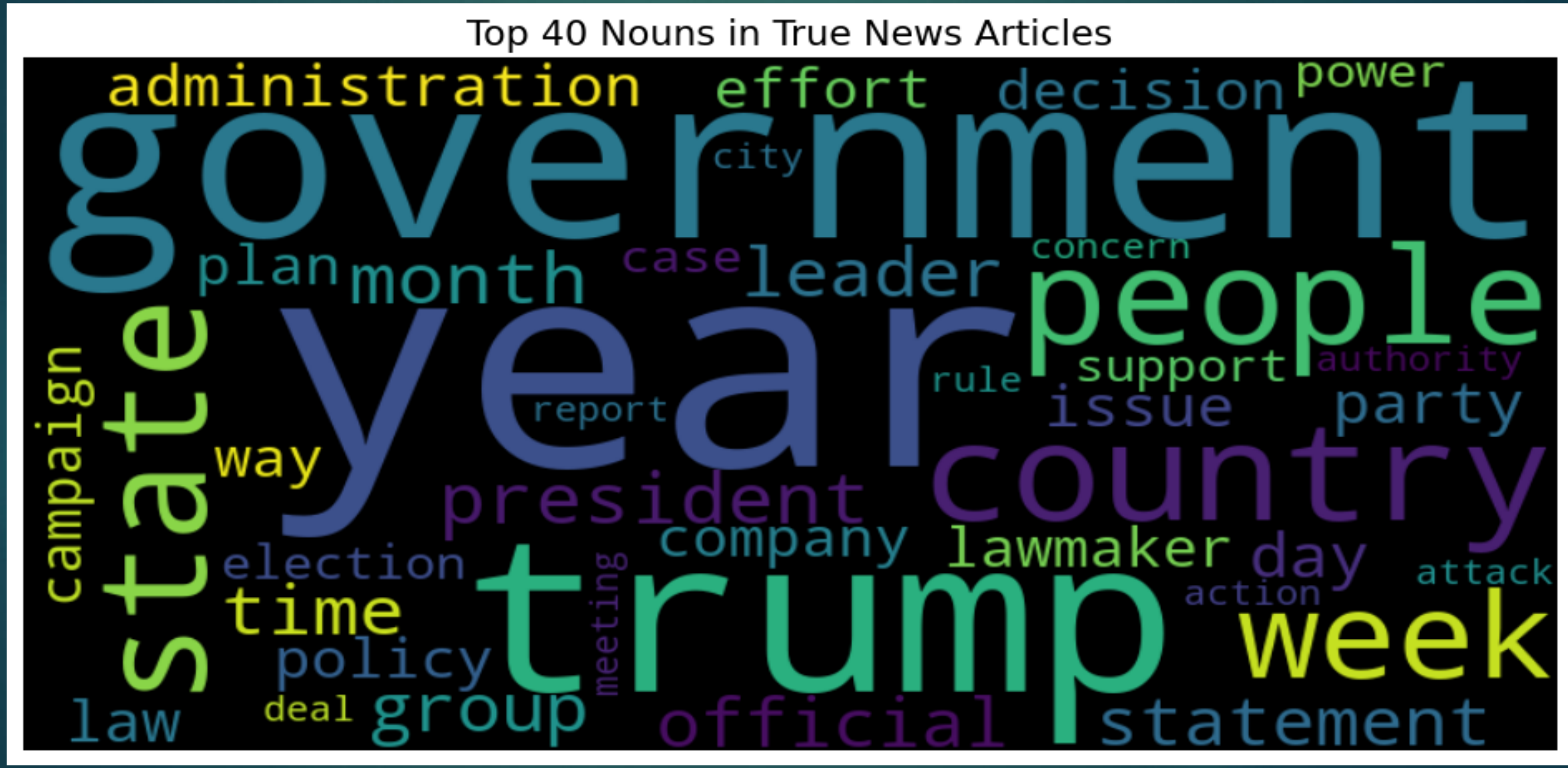
- **The Technical Goal:** Develop a semantic classification model that categorizes news based on **contextual meaning** rather than just syntax.

Understanding Content Structure



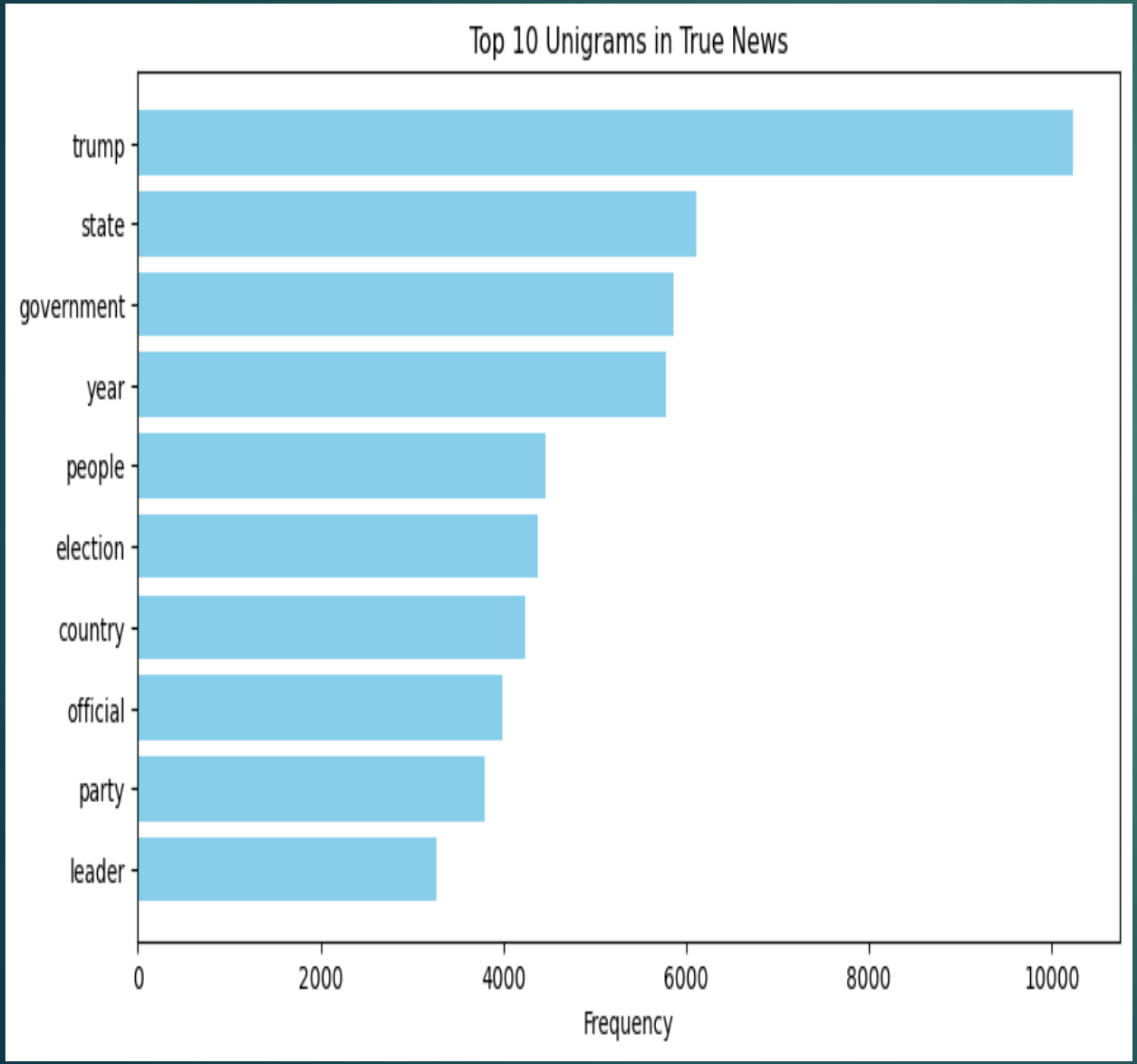
Cleaned news text typically follows a specific distribution; outliers in length can often be a signal for non-standard reporting or "clickbait."

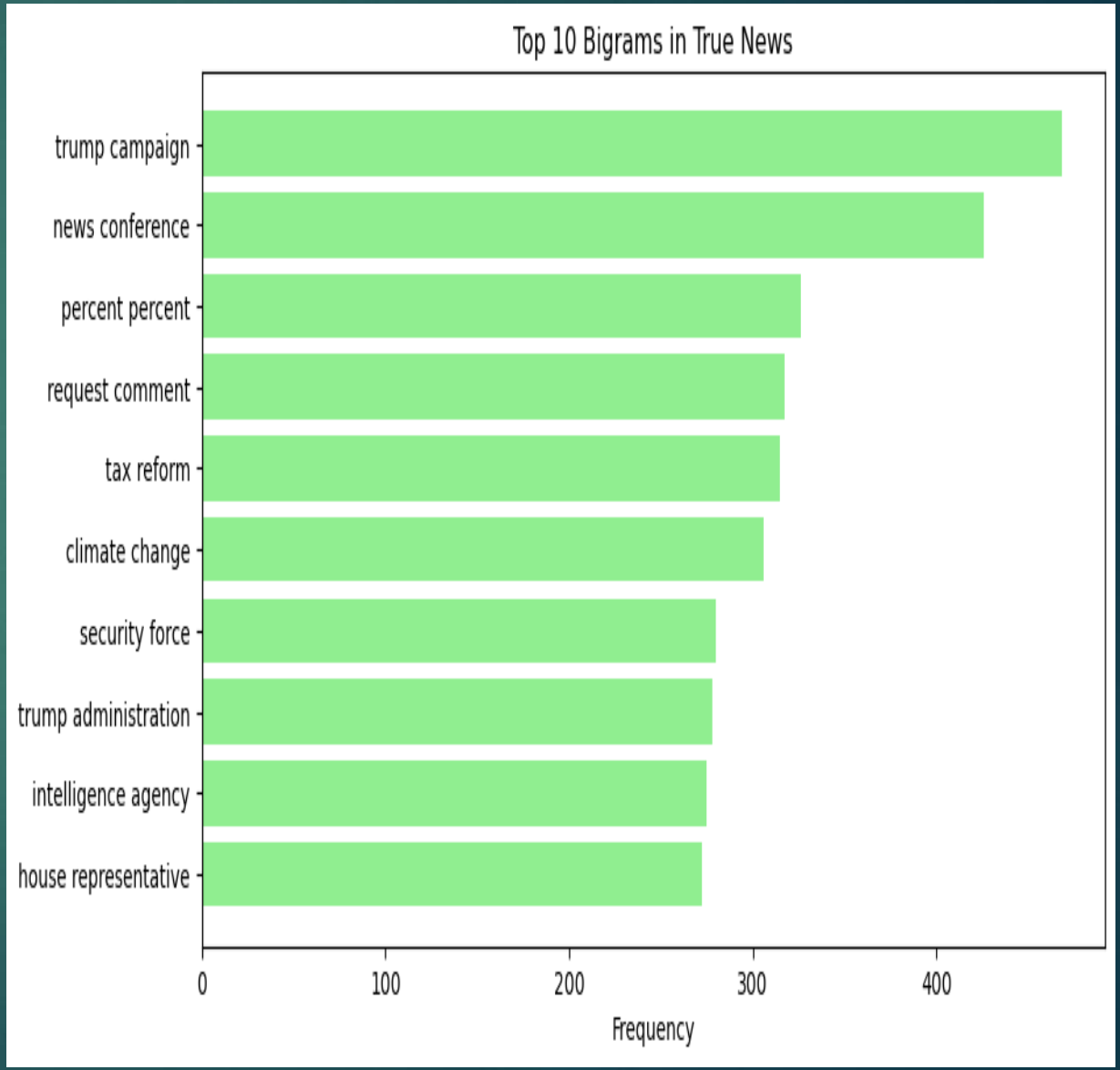
True News Analysis



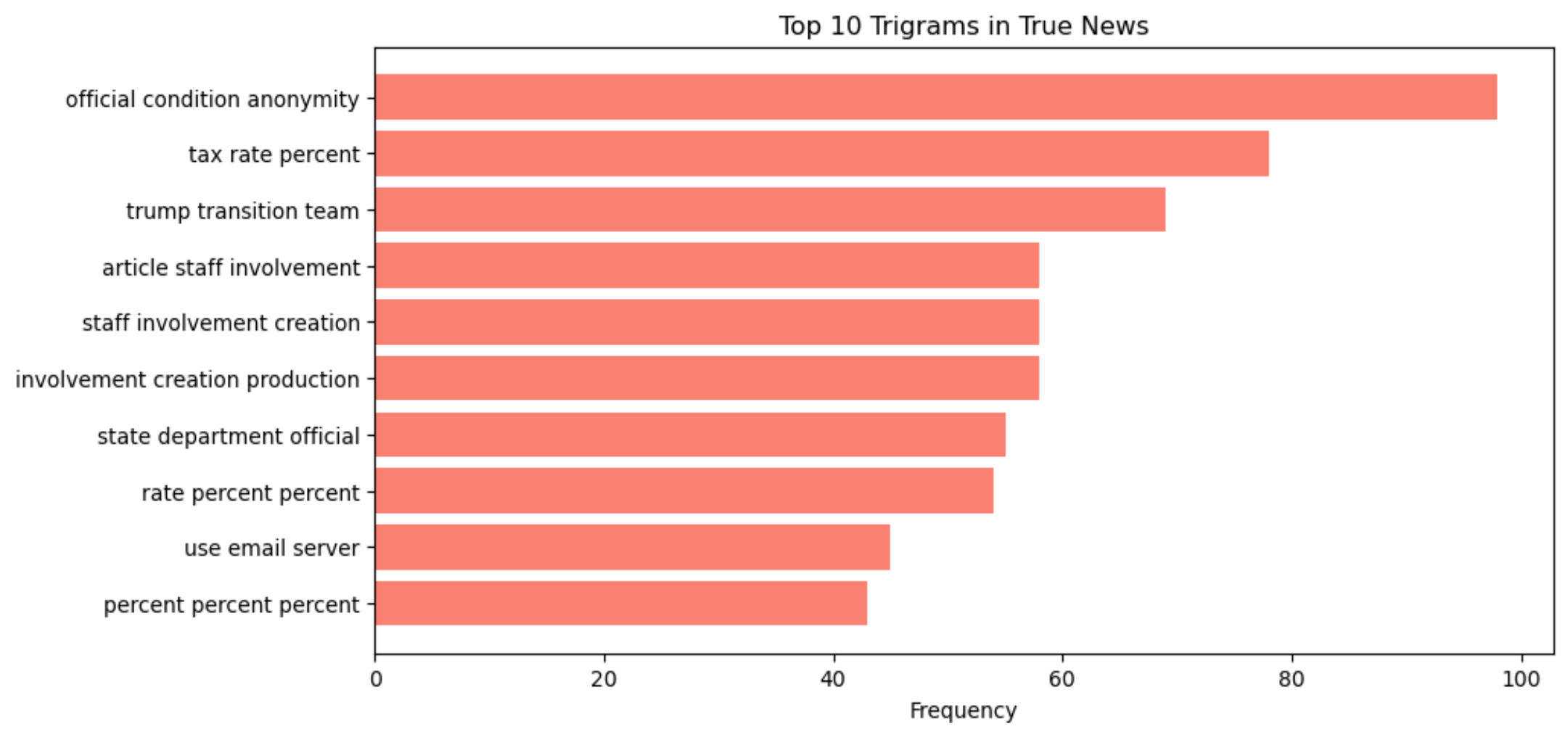
Fake News Analysis

The Language of Truth (N-Grams)





The Language of Truth (N-Grams)



Word2Vec Semantic Embeddings

- We utilized the **Google News 300** pre-trained model to convert text into 300-dimensional vectors.
- This allows the model to understand that "President" and "White House" are related, even if they aren't the same word.

Battle of the Algorithms

Logistic Regression Model	Decision Tree Model	Random Forest Model
Model Evaluation on Validation Data Accuracy: 0.9585 Precision: 0.9492 Recall: 0.9647 F1 Score: 0.9569	Decision Tree Evaluation on Validation Data Accuracy: 0.9027 Precision: 0.9079 Recall: 0.8858 F1 Score: 0.8967	Random Forest Evaluation on Validation Data Accuracy:0.9592 Precision:0.9572 Recall:0.9574 F1 Score:0.9573

Random Forest captures complex, non-linear patterns in the semantic space, making it the most robust choice for this task.

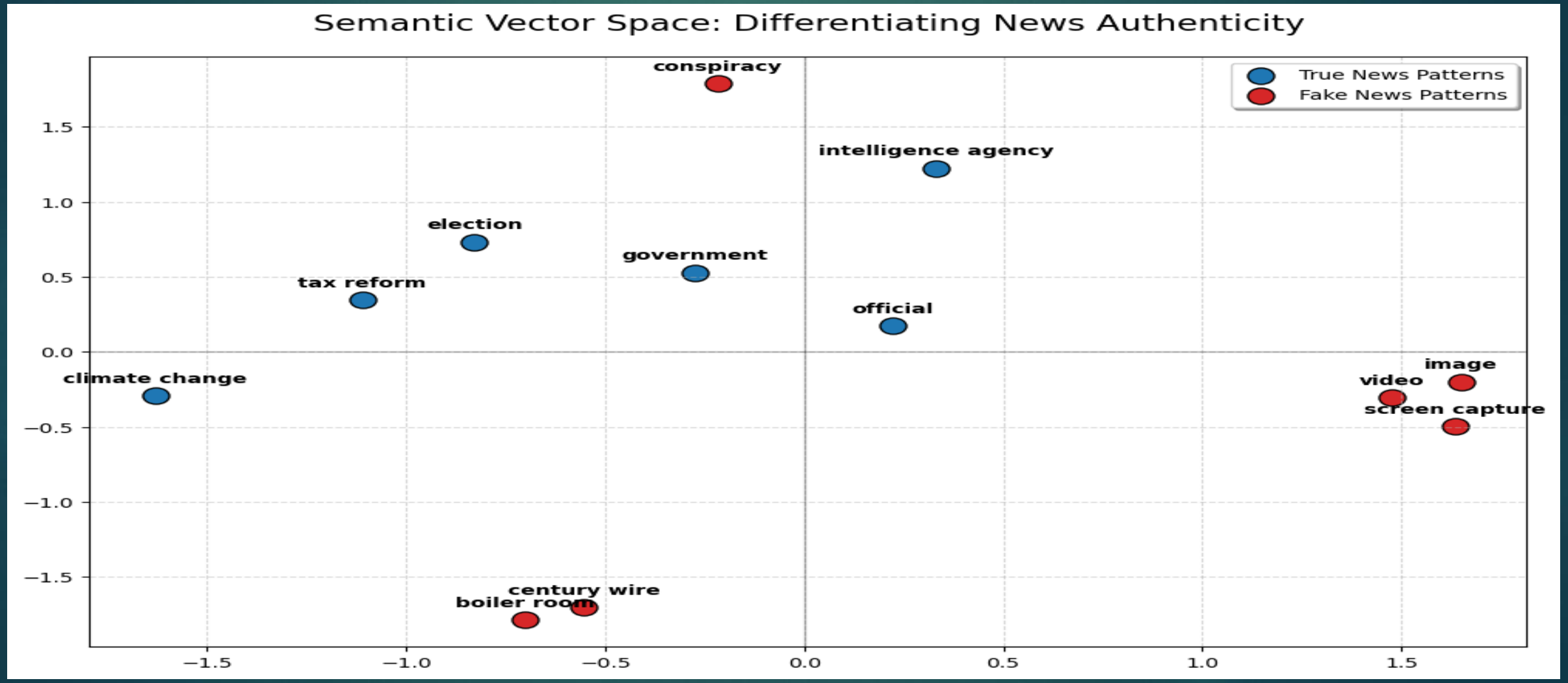
Deep Dive: Random Forest Performance

Classification Report:

	precision	recall	f1-score	support
Fake News	0.96	0.96	0.96	7045
True News	0.96	0.96	0.96	6425
accuracy			0.96	13470
macro avg	0.96	0.96	0.96	13470
weighted avg	0.96	0.96	0.96	13470

- The model achieves a high **F1-Score**, showing a healthy balance between Precision (not flagging real news as fake) and Recall (catching all fake news).
- Consistent performance across both classes indicates that the model is well-generalized and not biased.

Semantic Vector Space: Differentiating News Authenticity



Final Verdict: Scaling Truth through Semantic AI

Best Performing Model: The **Random Forest Classifier** emerged as the superior solution, chosen for its high **F1-Score**, which ensures a perfect balance between catching fake news (Recall) and not mislabeling credible news (Precision).

Semantic Advantage: Unlike traditional "word-matching" tools, our model successfully identified the **contextual difference** between factual, institutional reporting (True News) and sensationalist, emotionally charged narratives (Fake News).

Real-World Impact: This system provides a scalable, unbiased way to filter misinformation in real-time, reducing the manual effort of fact-checking and restoring trust in digital information ecosystems.