1. Objective

The objective of this project was to develop a **Semantic Classification model** that can automatically detect and classify news articles as either **true** or **fake**. This system aims to mitigate the spread of misinformation by analyzing the **meaning** of the text, rather than relying purely on syntax.

To achieve this, **Word2Vec** was used for semantic representation, and **supervised learning models** were trained to identify patterns in the text that differentiate between real and fake news articles.

2. Dataset Description

Two datasets were used:

- **True.csv** – 21,417 real news articles.
- **Fake.csv** – 23,502 fake news articles.

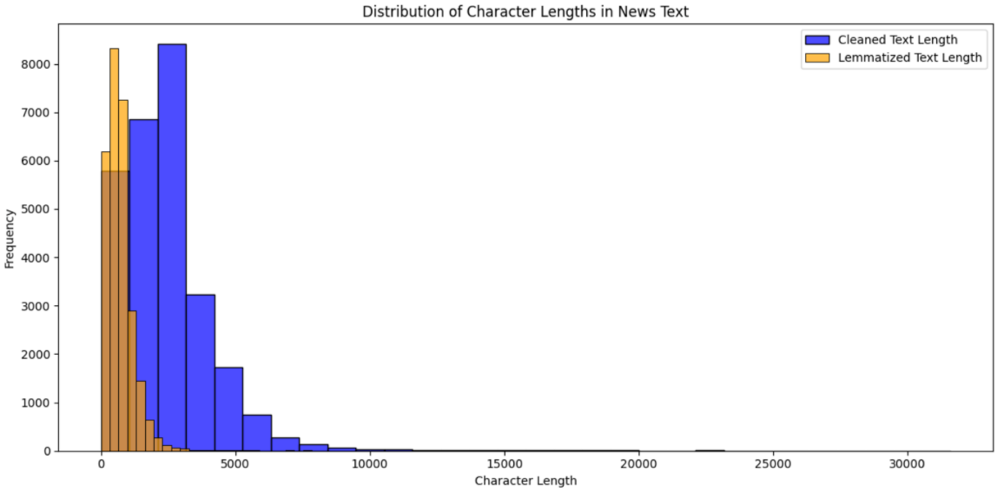
Each record included:

- Title of the news article
- Full text of the article
- Date of publication

3. Pipeline Overview

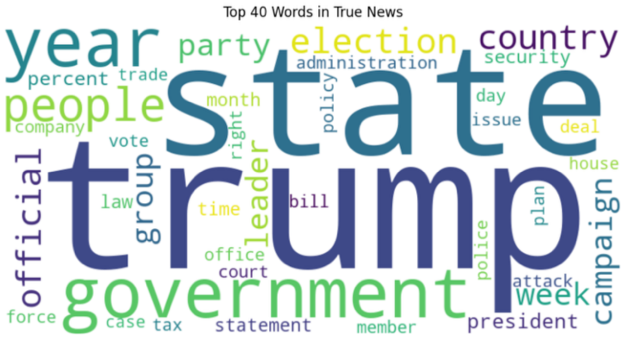
The following pipeline was executed:

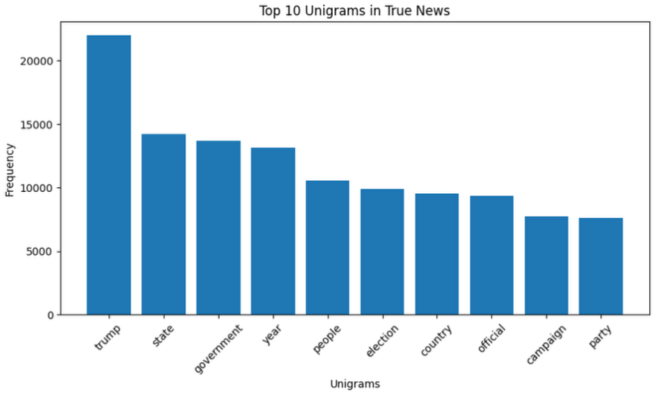
1. **Data Preparation**
2. **Text Preprocessing** (tokenization, stopword removal, lemmatization, etc.)
3. **Train-Validation Split** (ensuring balanced representation)
4. **Exploratory Data Analysis (EDA)**
    - **Top Unigrams, Bigrams, Trigrams** visualized using bar plots
    - **Word Clouds** created for both fake and real articles
5. **Feature Extraction** using **Word2Vec** for semantic representation
6. **Model Training and Evaluation** using three classifiers:
    - **Logistic Regression**
    - **Decision Tree**
    - **Random Forest**

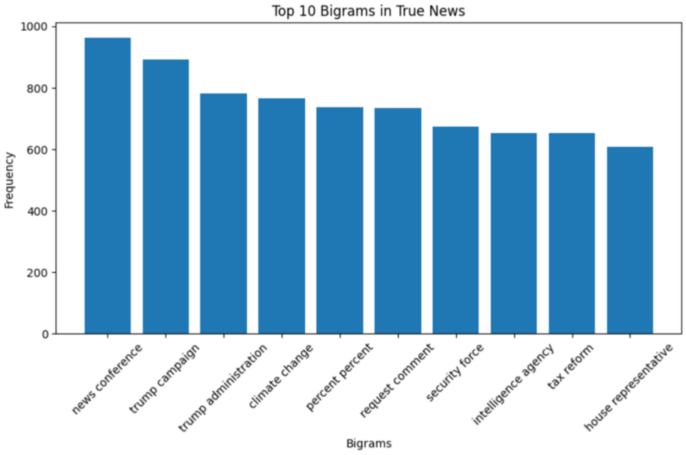


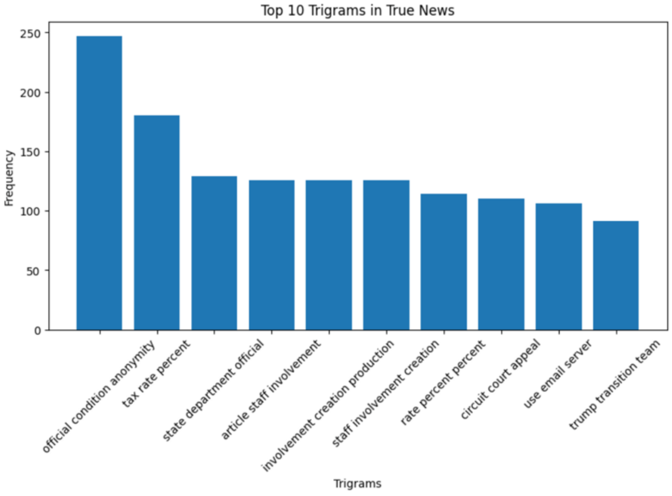
4. Key Observations from EDA

- **Real News Sources**:
  - Frequent terms related to **government policies**, **international affairs**, and **official announcements**.
  - More consistent language usage.
  - Unigrams Such as ‘trump’, ‘government’, and ‘state’ (Top 3 Results)
  - Bigrams such as ‘news conference’, ‘trump campaign’, and ‘trump administration’
  - Trigrams like ‘official condition anonymity’, ‘state department official’, and ‘tax rate percent’
  - Showed more sources and occurrence of official bodies such as the State and the Government.



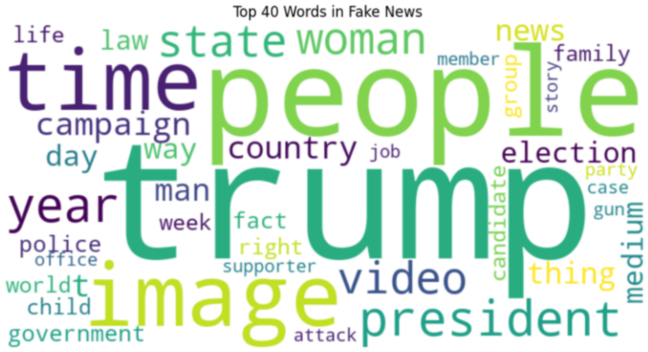


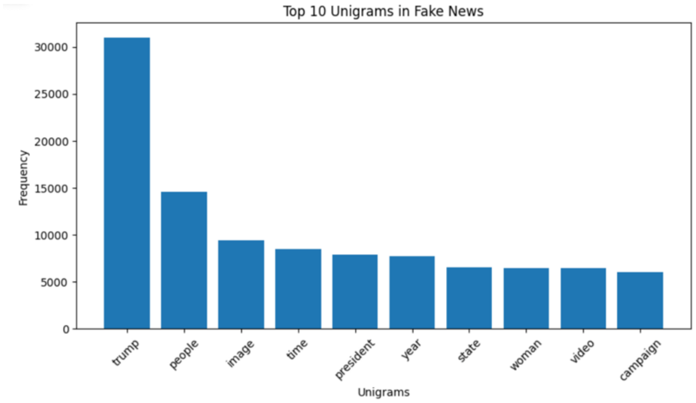


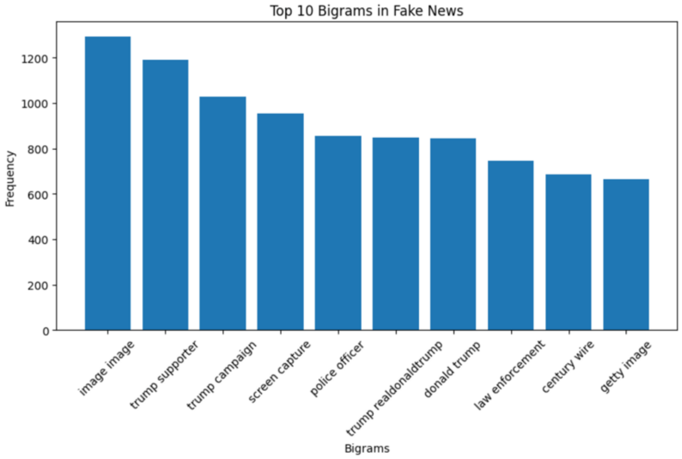


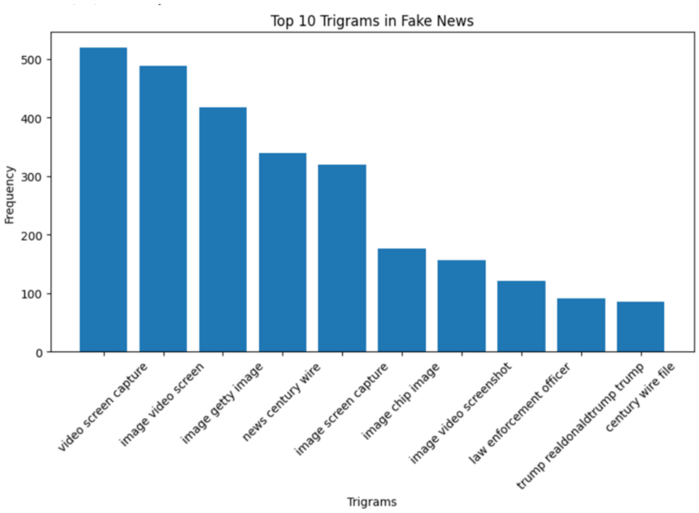
- **Fake News Sources**:
    - High frequency of emotionally charged or **conspiratorial language**.
    - Unigrams Such as ‘trump’, ‘people’, and ‘image’ (Top 3 Results)
    - Bigrams such as ‘image image’, ‘trump supporter’, and ‘trump campaign’

- Trigrams like ‘video screen capture’, ‘image video screen’, and ‘image getty image’
- Occurrence of sources such as twitter and youtube, which could be considered unofficial sources









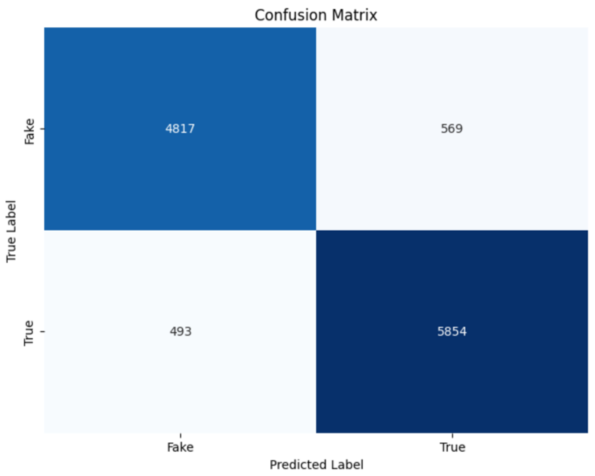
- Word clouds visually reinforced that **real articles used objective, formal vocabulary**, while **fake articles leaned towards sensationalist terms**.

5. Model Performance

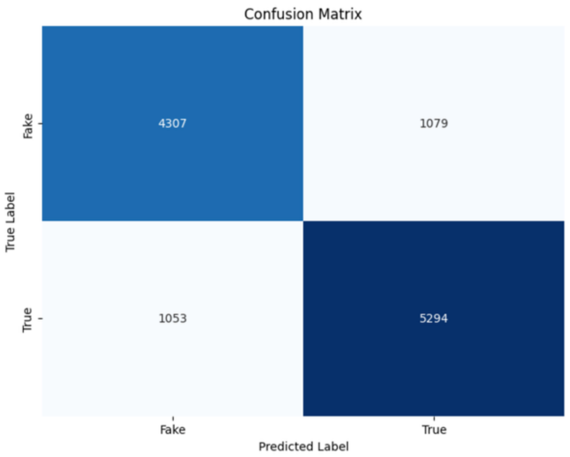
Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.9094**	**0.9114**	**0.9223**	**0.9168**
Decision Tree	0.8182	0.8306	0.8340	0.8323
Random Forest	0.8779	0.8717	0.9079	0.8894

Classification Report Insights:

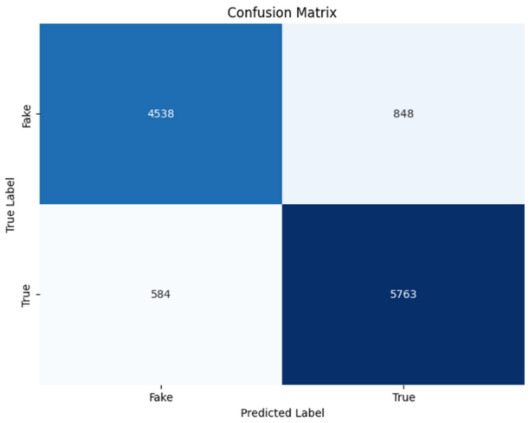
- **Logistic Regression** achieved the best overall balance of **precision, recall, and F1 score**, with slightly higher recall than Random Forest.
- **Random Forest** was a close second, performing robustly with slightly better precision.
- **Decision Tree**, while interpretable, underperformed in comparison to the other two models.
- The confusion matrix showed that the model had a low false positive and false negative rate, indicating its reliability.



Confusion Matrix for Logistic Regression



Confusion Matrix for Decision Tree Model



Confusion Matrix for Random Forest Model

6. Conclusion

The project successfully implemented semantic classification using **Word2Vec embeddings** to detect fake news, moving beyond keyword detection to understanding textual context and meaning.

Key patterns observed:

- **Fake news articles** used emotionally manipulative and clickbait language.
- **Real news articles** maintained formal and structured language focused on facts and official statements.

Among the three models evaluated, **Logistic Regression** was selected as the best-performing model due to its **highest F1 score (0.9168)** and **balanced precision-recall performance**. This metric was prioritized to maintain a trade-off between detecting fake articles (recall) and ensuring accuracy in predictions (precision).

The semantic approach allowed the system to not only identify superficial word patterns but also **understand contextual relevance**, improving robustness against simple lexical tricks often used in fake content.