- ❖ **Title**: Semantic Classification Assignment – **Fake News Detection**
- ❖ **Name**: Arjun Deepak

Problem Statement & Objective:

The objective of this assignment is to develop a Semantic Classification model. You will be using Word2Vec method to extract the semantic relations from the text and develop a basic understanding of how to train supervised models to categorise text based on its meaning, rather than just syntax. You will explore how this technique is used in situations where understanding textual meaning plays a critical role in making accurate and efficient decisions.

Business Objective:

The spread of fake news has become a significant challenge in today's digital world. With the massive volume of news articles published daily, it's becoming harder to distinguish between credible and misleading information. This creates a need for systems that can automatically classify news articles as true or fake, helping to reduce misinformation and protect public trust.

In this assignment, you will develop a Semantic Classification model that uses the Word2Vec method to detect recurring patterns and themes in news articles. Using supervised learning models, the goal is to build a system that classifies news articles as either fake or true.

Pipelines that need to be performed

You need to perform the following tasks to complete the assignment:

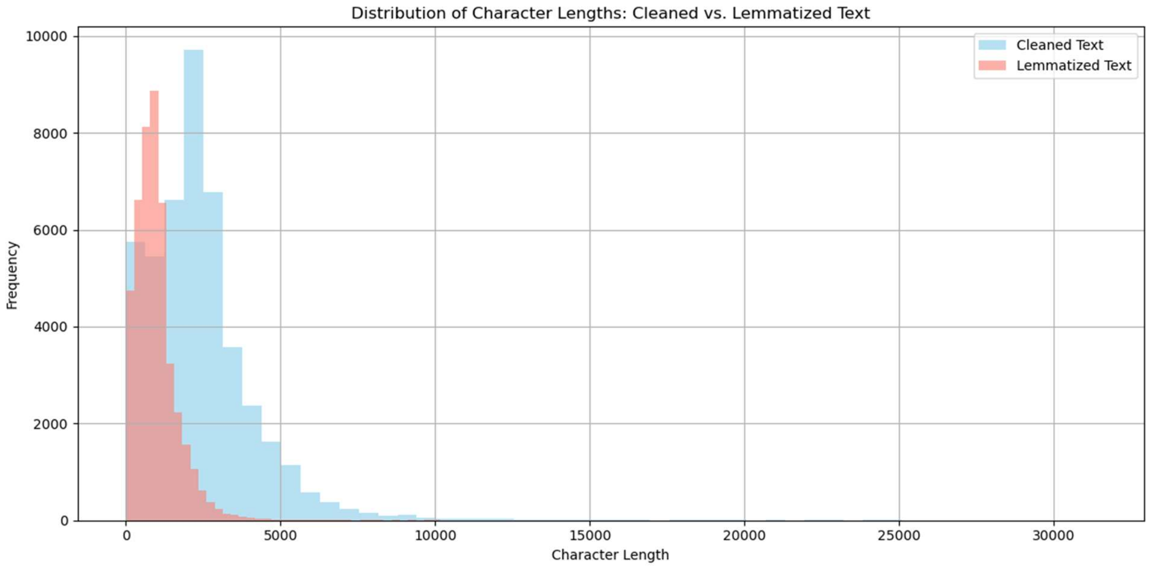
- ✓ Data Preparation
- ✓ Text Preprocessing
- ✓ Train Validation Split
- ✓ EDA on Training Data
- ✓ EDA on Validation Data [Optional]
- ✓ Feature Extraction
- ✓ Model Training and Evaluation.

✓ Data Preparation

Summary:

The dataset was loaded and inspected. Null values were checked, and the class distribution between fake and true news articles was examined.

✓ Text Preprocessing



Plot 1: Histogram of Text Lengths

Explanation:

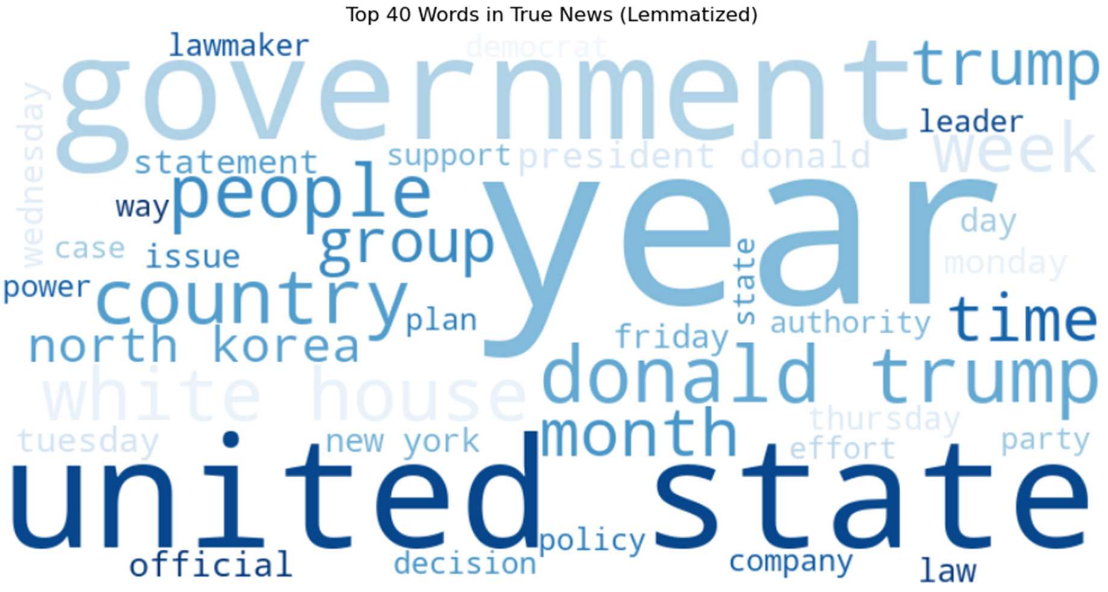
This histogram shows the number of characters in original, cleaned, and lemmatized texts. Cleaning reduced noise (e.g., punctuation, HTML tags), and lemmatization condensed texts by keeping only meaningful root words.

✓ Train-Validation Split

Summary:

The cleaned data was split into training and validation sets using an 70/30 ratio, preserving label distribution.

✓ EDA on Training Data



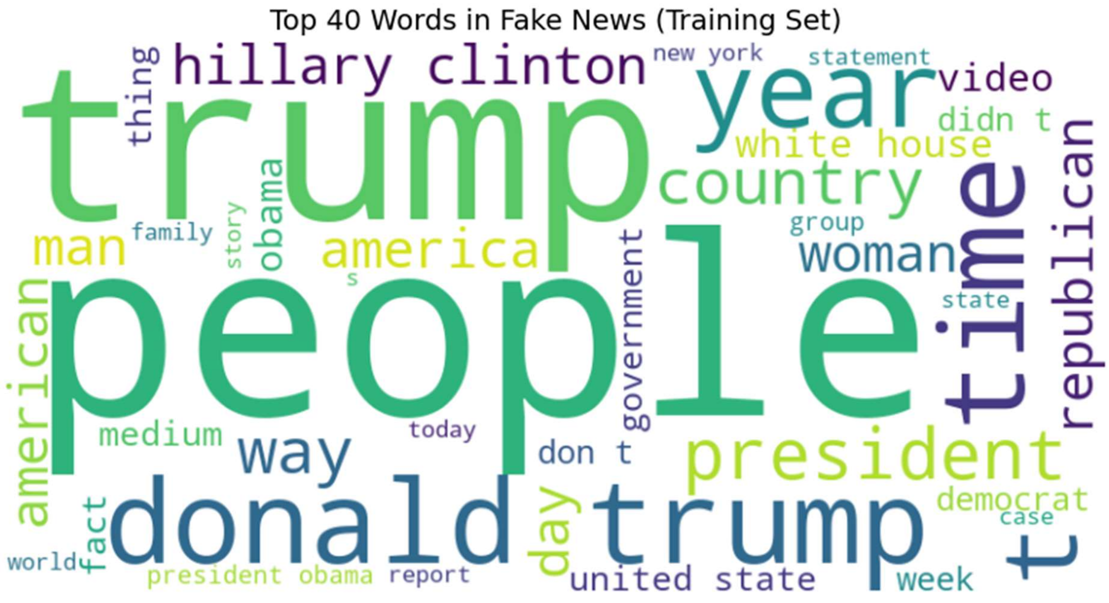
Plot 2: Word Cloud – True News

Explanation:

Frequent terms in true news show neutral, factual reporting such as “government,” “officials,” and “report.” The tone is formal and informative.

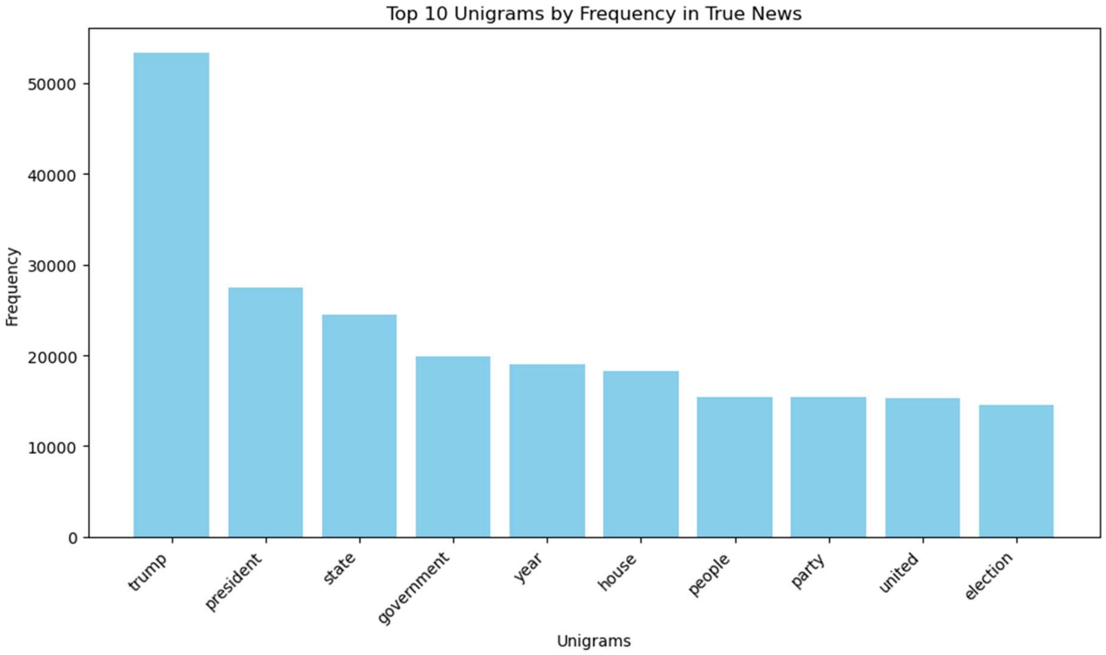
- ✓ **EDA on Training Data**

Word Cloud – Fake News:

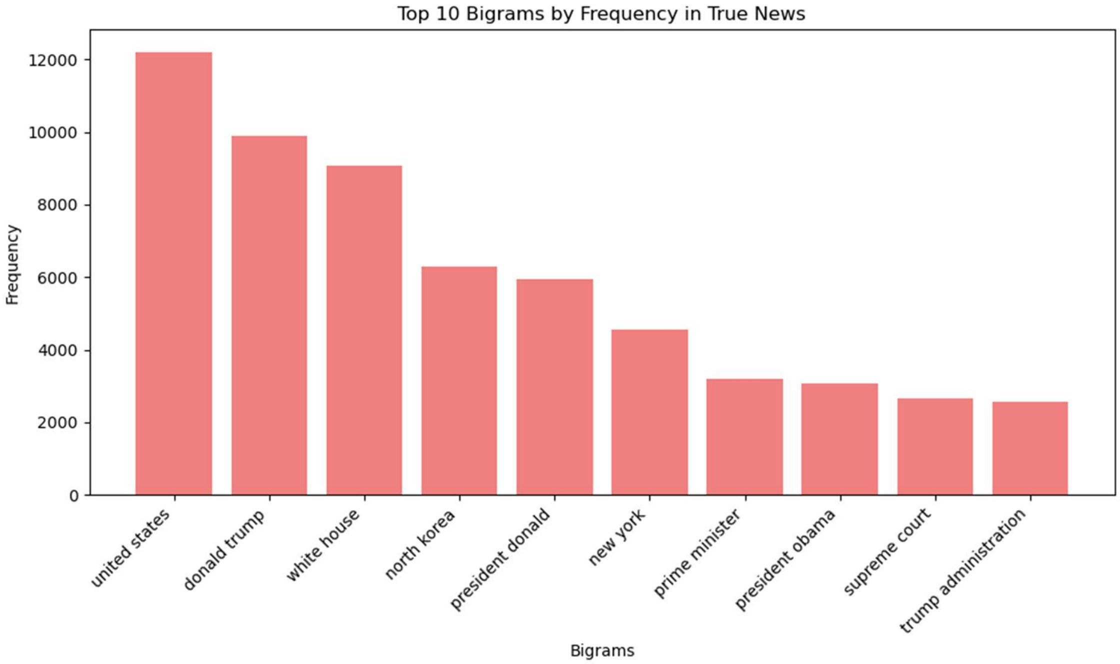


Explanation:

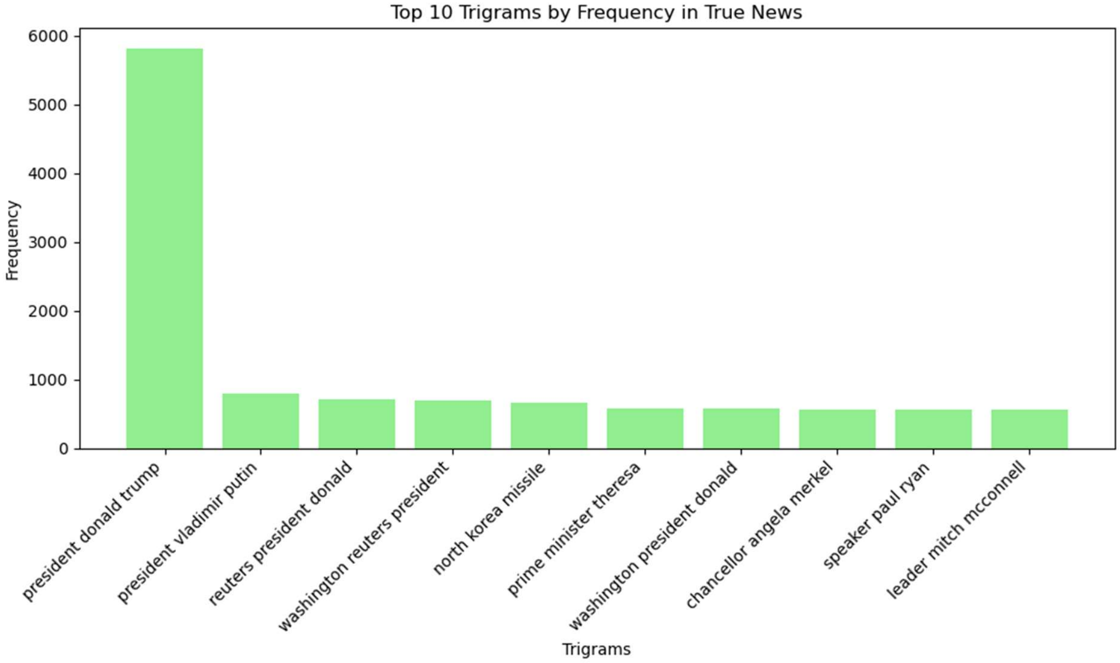
Fake news content is often dramatic or manipulative, with terms like “shocking,” “truth,” and “secret” suggesting sensationalism.



Top 10 Unigrams – True News



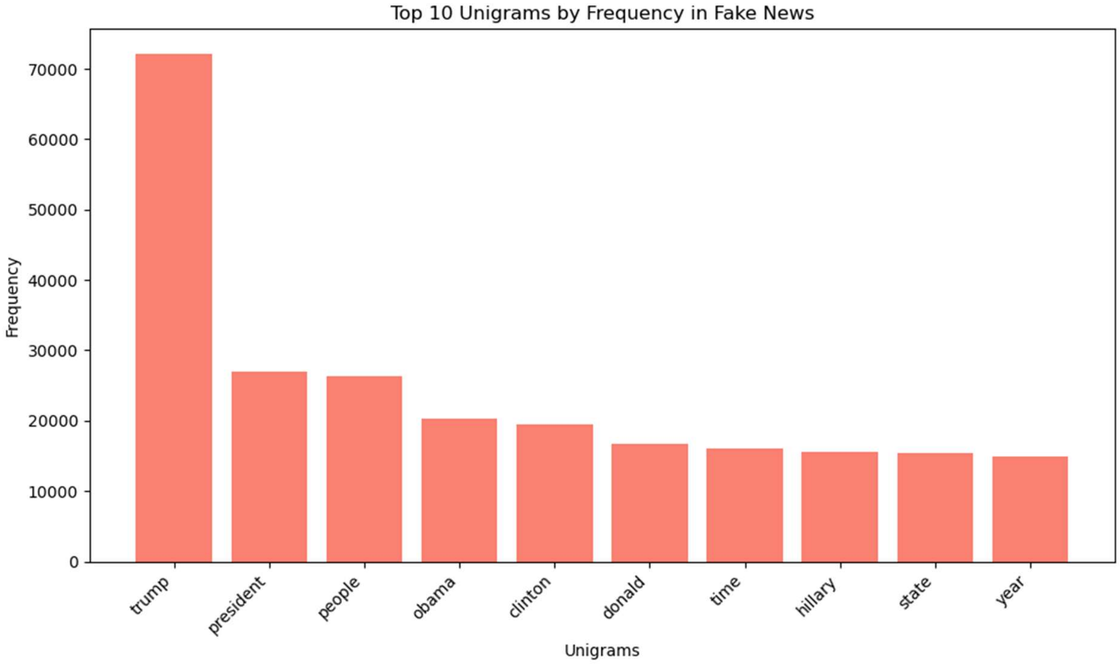
Top 10 Bigrams – True News

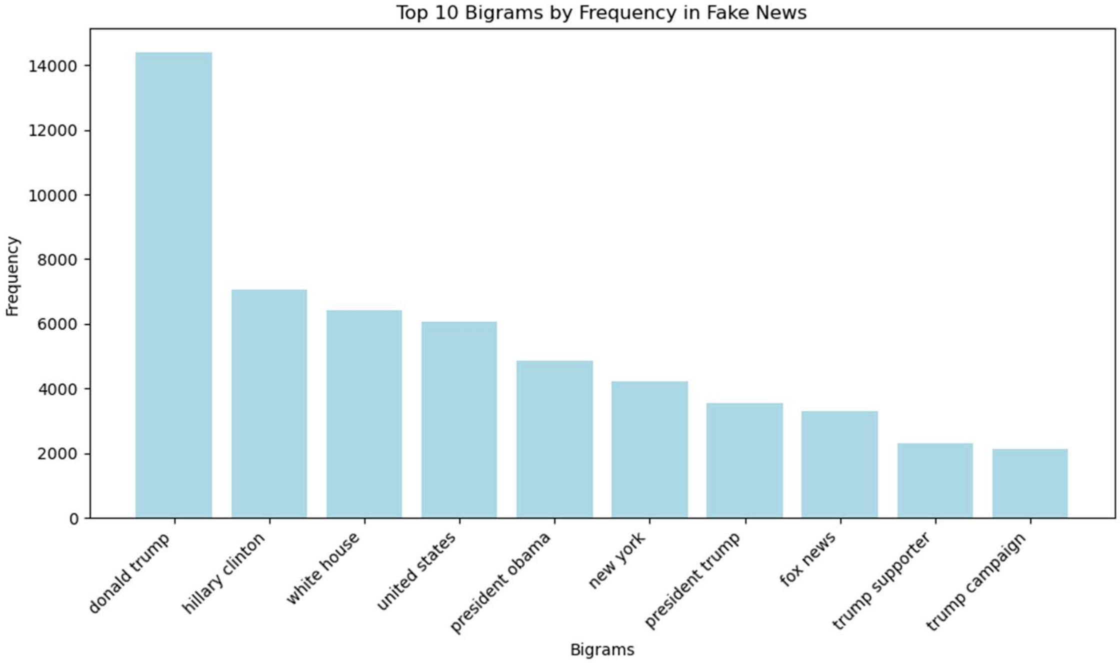


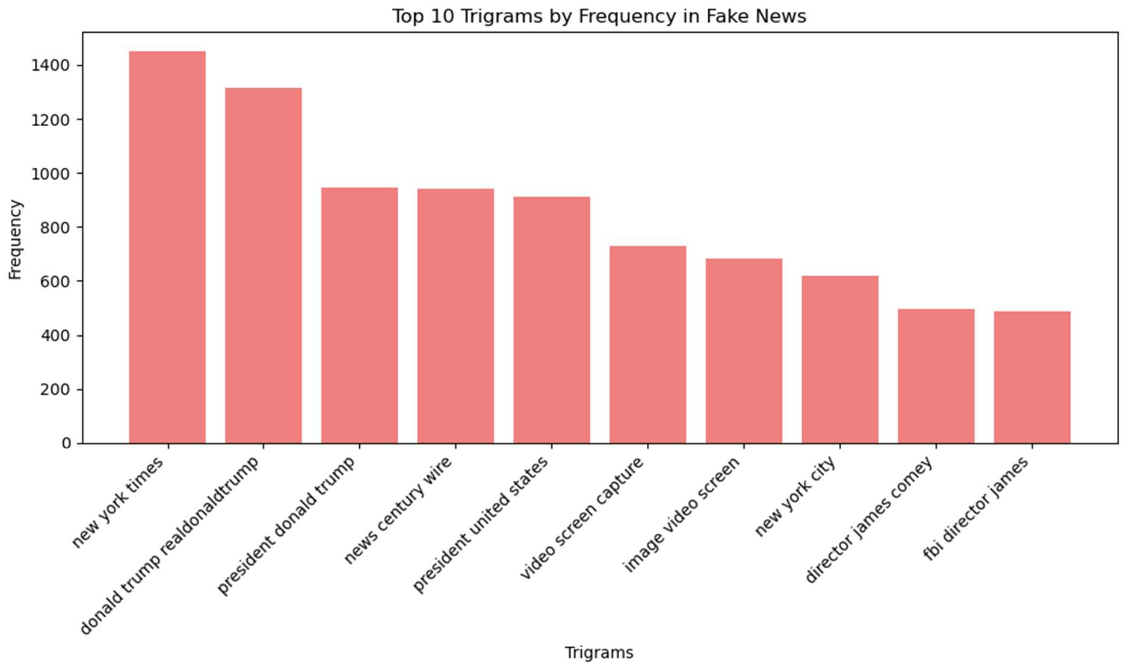
Top 10 Trigrams – True News

Explanation:

Unigrams, Bigrams and trigrams in true news reflect objective language and source citation patterns like “according to reports.”







Explanation:

These n-grams reveal repetition of names, exaggerated terms, or conspiratorial phrases. The difference in tone is a useful signal for classification.

✓ Feature Extraction

Explanation:

Word2Vec embeddings were used to convert text into dense vector representations, capturing semantic meaning of words. The average word vectors were taken for each text document.

✓ Model Training and Evaluation

Explanation:

Among the models trained:

- Logistic Regression achieved the highest **F1-score (0.9313)** and accuracy (**93.39%**),
- Random Forest followed closely,
- Decision Tree underperformed comparatively.

Logistic Regression Model:

Accuracy: **0.9339**

Precision: 0.9227

Recall: 0.9401

F1 Score: **0.9313**

Decision Tree Model:

Accuracy: 0.8513

Precision: 0.8563

Recall: 0.8271

F1-score: 0.8414

Random Forest Model:

Accuracy: 0.9291

Precision: 0.9317

Recall: 0.9188

F1-score: 0.9252

✓ Conclusion & Insights

Through semantic classification using Word2Vec and traditional ML classifiers, we successfully differentiated true and fake news articles. Analysis of linguistic patterns showed fake news uses emotionally charged or vague language, while true news aligns with fact-based reporting.

The **Logistic Regression model** delivered the best performance with an **accuracy of 93.39%**, making it the most effective classifier in this task. F1 score was prioritized as the key metric due to its balance between precision and recall.

This approach demonstrates that semantic analysis, combined with classical machine learning, can be a powerful method for tackling misinformation.