Fake News Detection

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.

Data Dictionary

There are two datasets, True.csv and Fake.csv. Both datasets contain three columns:

- title of the news article
- text of the news article
- date of article publication

True.csv dataset includes 21,417 true news, while the Fake.csv dataset comprises 23,502 fake news.

Patterns Observed in True and Fake News:

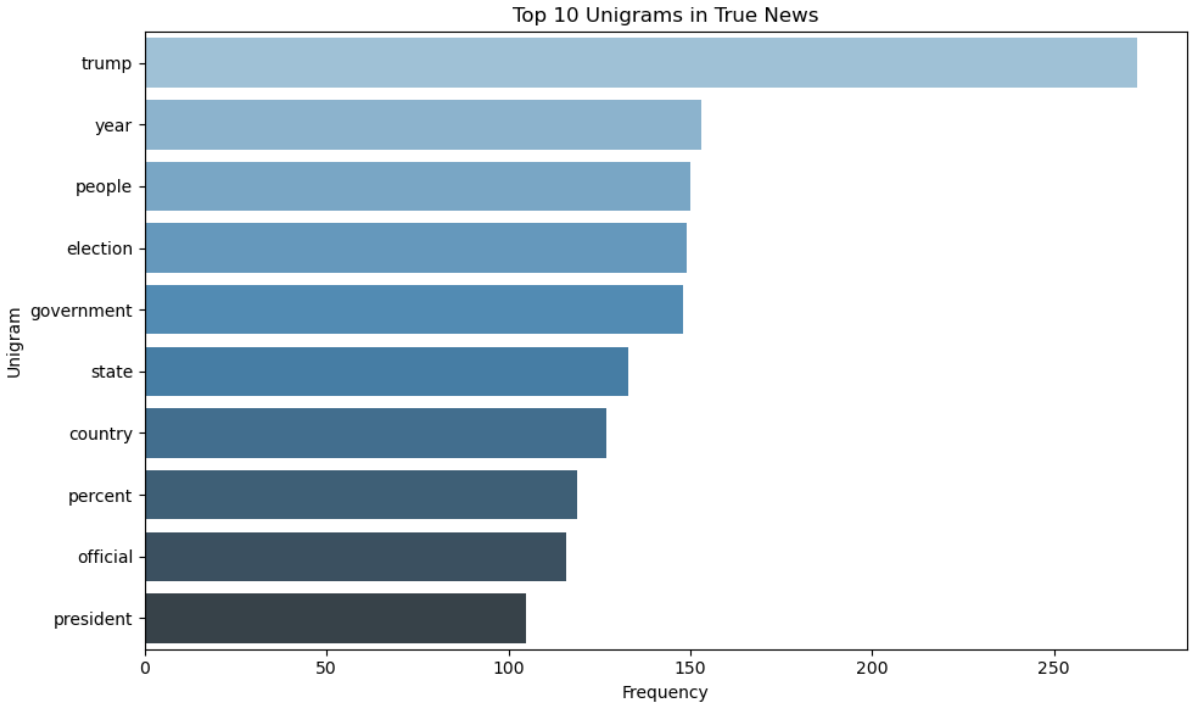
- **True news** articles tend to use more formal language, factual nouns, and domain-specific terminology. Their text exhibits consistent patterns with reliable information markers.

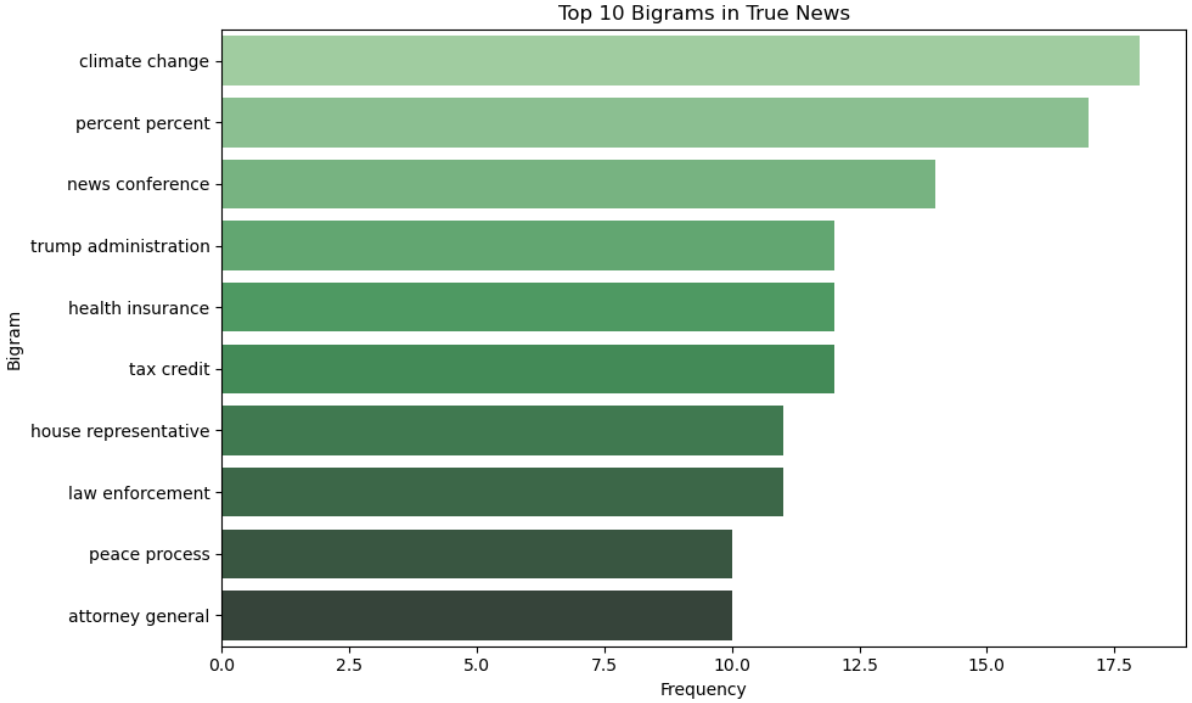
Top 40 Words in True News (Training Data)

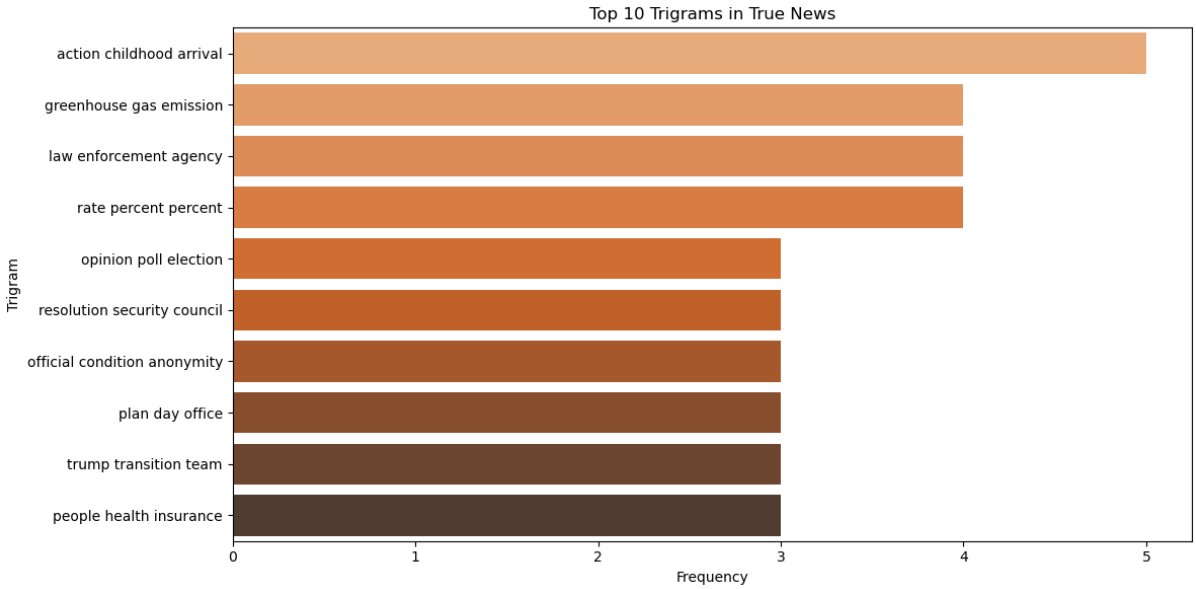
- **Fake news** often contains sensational or emotionally charged language, repetitive phrases, and sometimes less coherent structure. The linguistic style differs noticeably, with certain keywords and n-grams appearing frequently that indicate misinformation or bias.

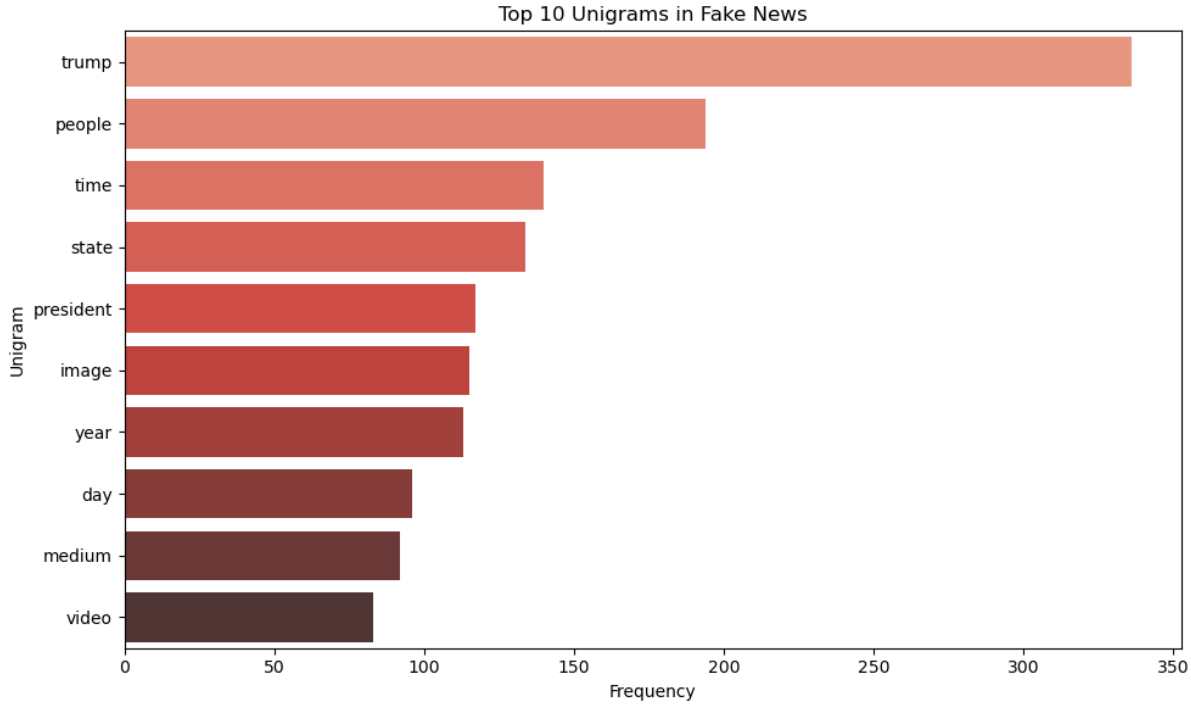
Top 40 Words in Fake News (Training Data)

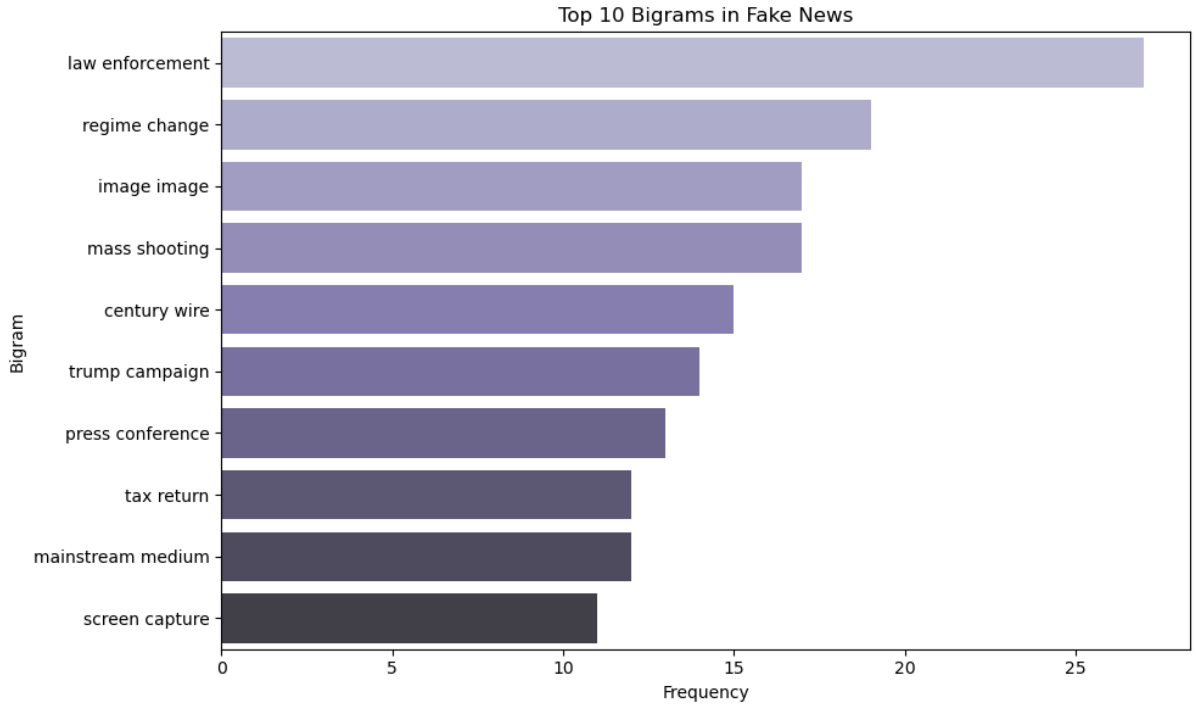
- Word frequency analysis (unigrams, bigrams, trigrams) and word clouds highlighted distinct vocabularies for true vs. fake news, supporting these linguistic differences.

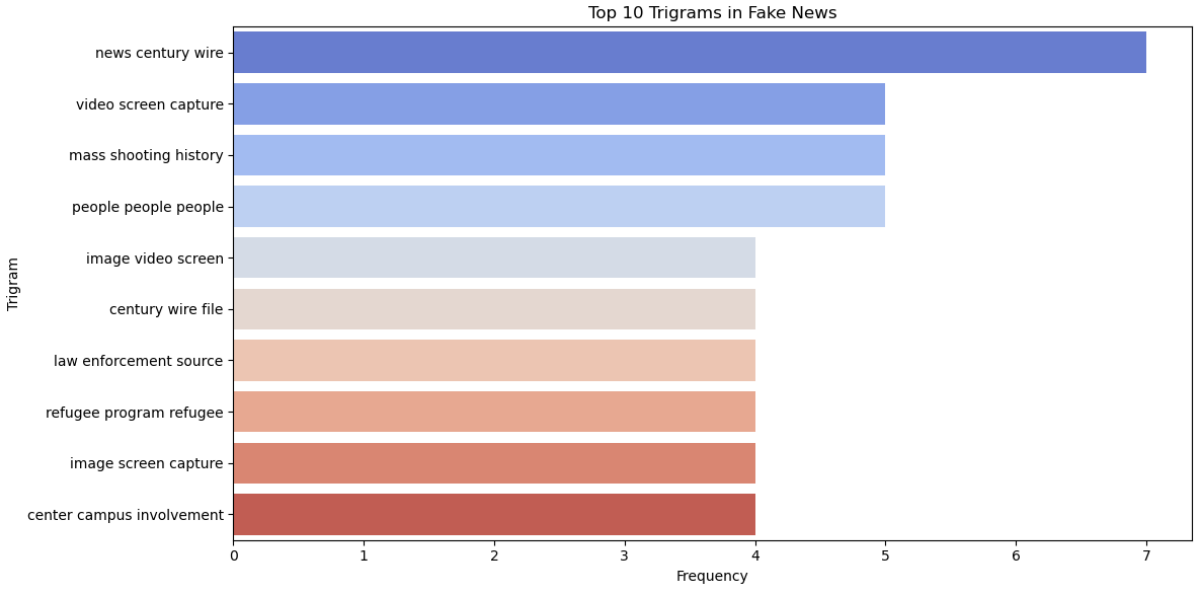

Top 10 Unigrams in True News
trump
year
people
election
government
state
country
percent
official
president
Unigram
0
50
100
150
200
250
Frequency


Top 10 Bigrams in True News
climate change
percent percent
news conference
trump administration
health insurance
tax credit
house representative
law enforcement
peace process
attorney general
Bigram
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Frequency


Top 10 Trigrams in True News
action childhood arrival
greenhouse gas emission
law enforcement agency
rate percent percent
opinion poll election
resolution security council
official condition anonymity
plan day office
trump transition team
people health insurance
Trigram
0
1
2
3
4
5
Frequency

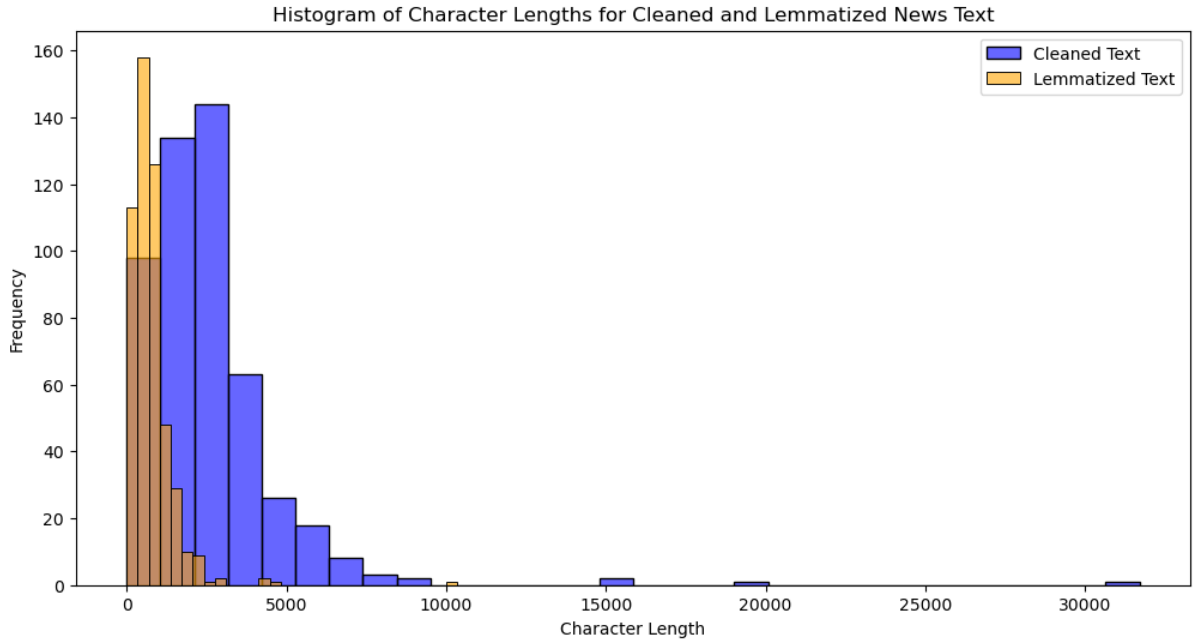

Top 10 Unigrams in Fake News
trump
people
time
state
president
image
year
day
medium
video
Unigram
0
50
100
150
200
250
300
350
Frequency





Semantic Classification Approach:

- The approach combined text cleaning, lemmatization, and part-of-speech filtering (focusing on nouns) to capture meaningful semantic features.



- Pre-trained Word2Vec embeddings were used to transform news text into dense vector representations, effectively capturing semantic relationships beyond simple word counts.
- These embeddings enabled classical machine learning classifiers (Logistic Regression, Decision Tree, Random Forest) to learn subtle patterns in word usage for distinguishing true and fake news.

Model Performance and Selection:

- Among the models trained, **Random Forest** consistently achieved the best balance of accuracy, precision, recall, and F1-score on validation data.
- The **F1-score** was prioritized as the key evaluation metric because it balances false positives and false negatives, crucial for minimizing both misinformation spread and wrongful censorship.
- Logistic Regression showed good baseline performance, but Random Forest's ability to model complex feature interactions led to superior results.
- Decision Tree, while interpretable, underperformed compared to ensemble methods.

Assessment and Impact:

- The semantic classification pipeline demonstrated the effectiveness of combining linguistic preprocessing with advanced embeddings for fake news detection.
- By focusing on nouns and semantic content, the model reduced noise from irrelevant words, improving classification robustness.
- The use of pre-trained Word2Vec allowed leveraging rich semantic knowledge without extensive training data or deep learning architectures.
- This approach offers a practical, scalable solution for automated news verification, potentially aiding platforms in combating misinformation.

- Future improvements could include fine-tuning embeddings on domain-specific data or integrating transformer-based contextual embeddings for even better performance.