

Fake News Detection Using Word2Vec and Machine Learning

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.

Steps performed

Data Preparation

Text Preprocessing

Train Validation Split

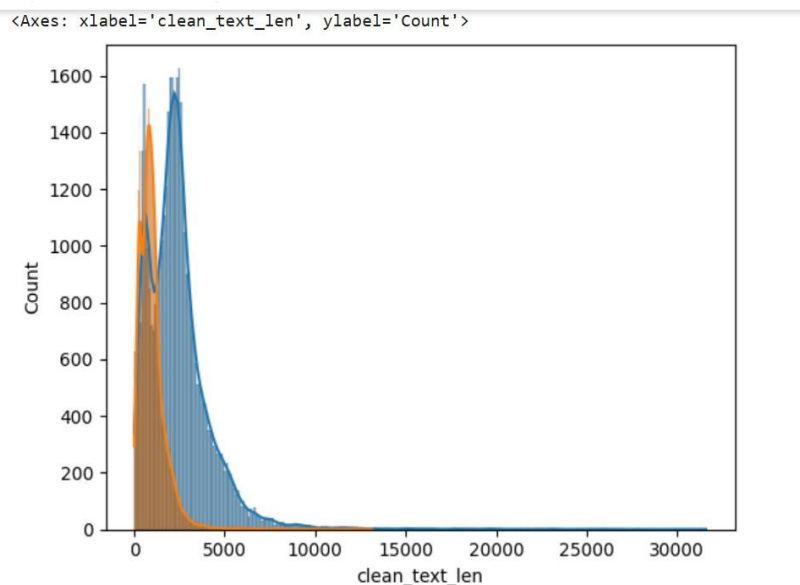
EDA on Training Data

Feature Extraction

Model Training and Evaluation

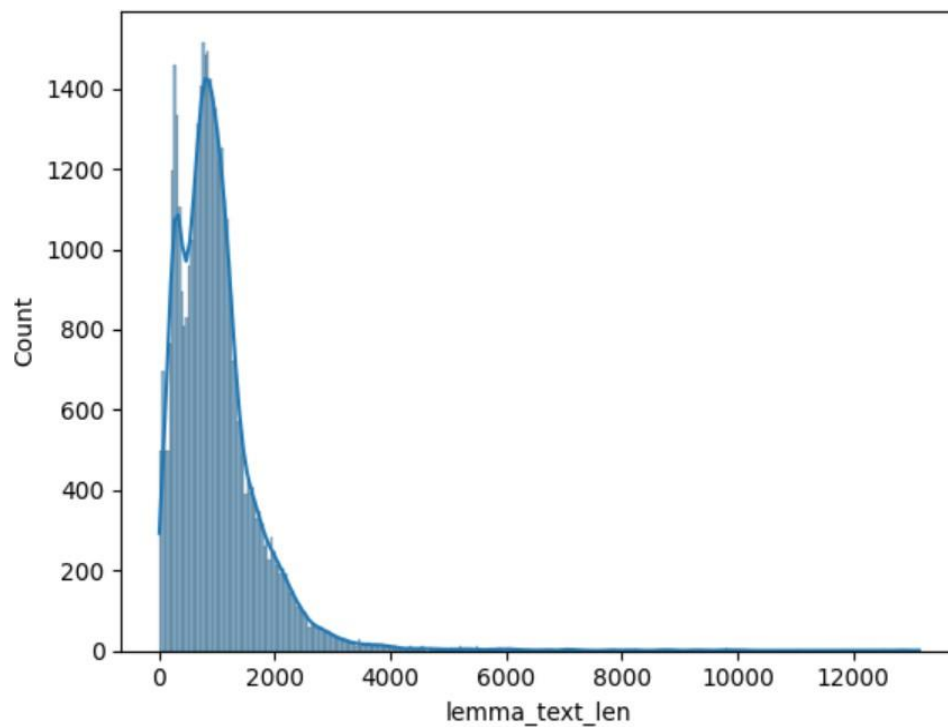
Visualizations

Histogram with cleaned news text and lemmatized news text



This histogram shows the distribution of text lengths in the cleaned news dataset. It is right-skewed.

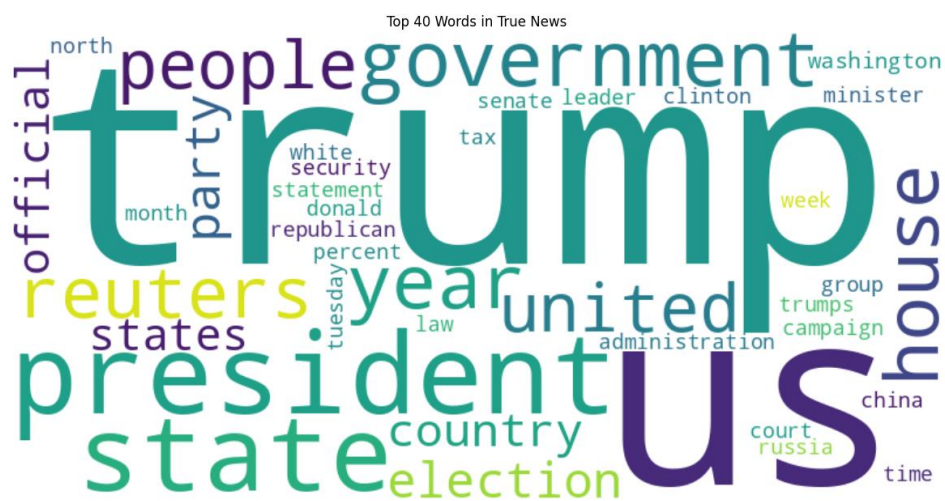
Histogram for lemmatized news text with POS tags removed



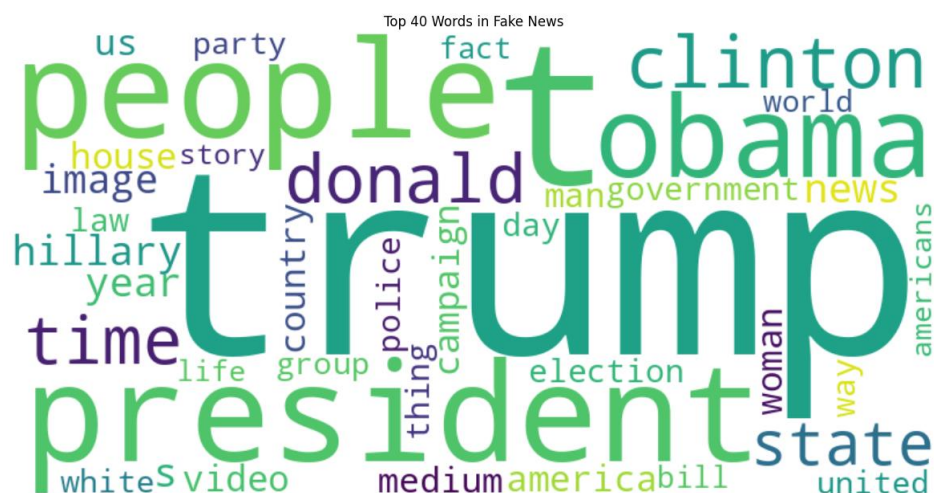
Word Frequency Analysis

Most occurred 40 words in True news and Fake News

True News: [trump, us, president, state, government, year, house, reuters, people, united, election, party, country, official, states, washington, group, campaign, leader, court, security, week, donald, minister, percent, clinton, senate, law, white, north, month, republican, administration, china, statement, time, russia, trumps, tax, tuesday]



Fake News: [trump, t, president, people, obama, clinton, donald, time, state, hillary, s, year, news, image, video, woman, us, campaign, house, america, country, medium, man, election, day, way, thing, government, police, law, party, white, group, united, fact, world, bill, story, life, americans]



Bar Plot of top 10 unigrams by frequency in true news and fake news

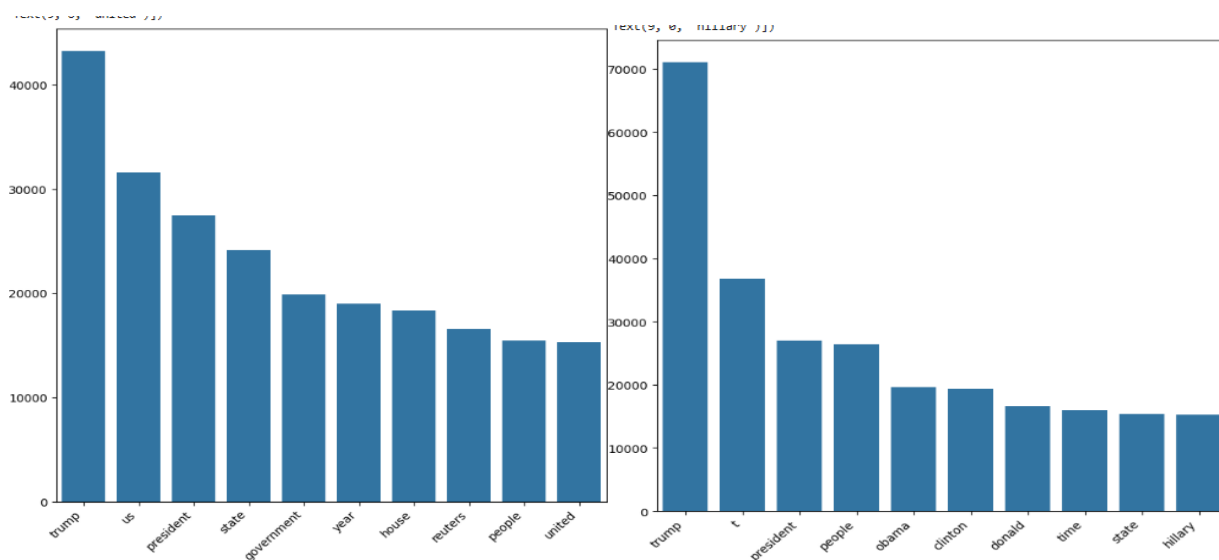


Fig: Bar Plots of top 10 unigrams in true news Fig: Bar Plots of top 10 unigrams in fake news

Bar Plot of top 10 bigrams by frequency in true news and fake news

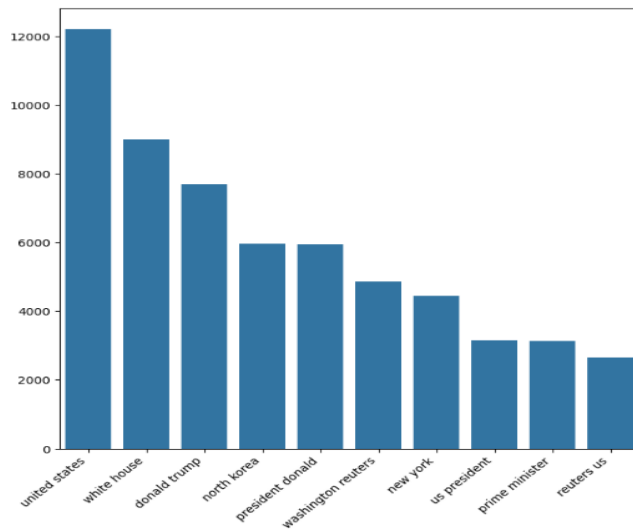


Fig: Bar Plots of top 10 bigrams in true news

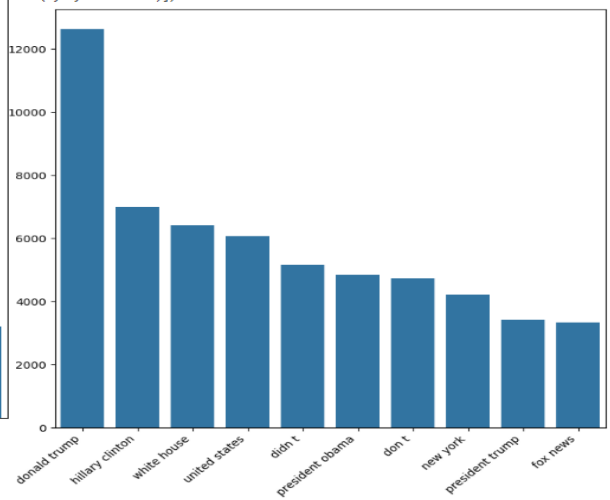


Fig: Bar Plots of top 10 bigrams in fake news

Bar Plot of top 10 trigrams by frequency in true news

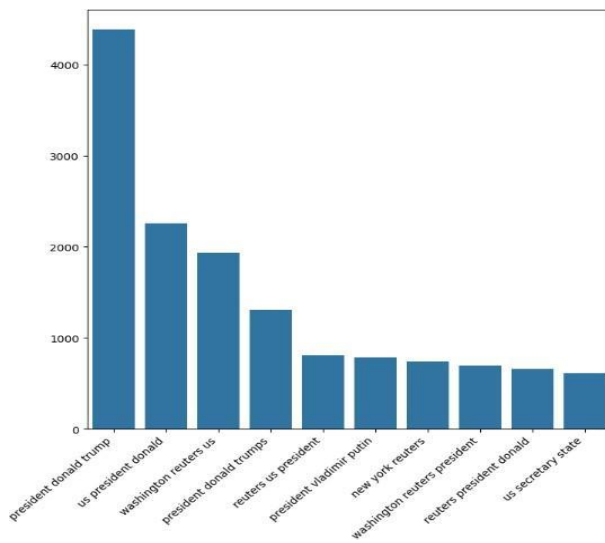


Fig: Bar Plots of top 10 trigrams in true news

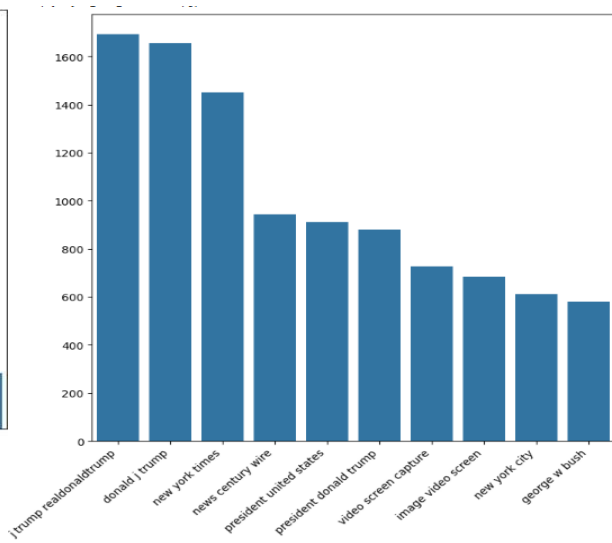


Fig: Bar Plots of top 10 trigrams in fake news

Results [Evaluation Metrics and Confusion Matrix]

Confusion Matrix– Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
0 (True News)	0.90	0.82	0.86	7780
1 (Fake News)	0.78	0.88	0.83	5690
Accuracy			0.84	13470
Macro Avg	0.84	0.85	0.84	13470
Weighted Avg	0.85	0.84	0.84	13470

- The model achieves an **overall accuracy of 84%**, indicating strong performance.
- It performs **slightly better on class 0 (True News)** in terms of precision.
- **Recall is higher for class 1 (Fake News)**, meaning the model captures fake news more effectively than it misses.
- Balanced F1-scores across classes show a good trade-off between precision and recall.

Confusion Matrix– Decision Tree Model

Class	Precision	Recall	F1-Score	Support
0 (True News)	0.88	0.86	0.87	7171
1 (Fake News)	0.85	0.86	0.85	6299
Accuracy			0.86	13470
Macro Avg	0.86	0.86	0.86	13470
Weighted Avg	0.86	0.86	0.86	13470

- The **Decision Tree classifier achieved an accuracy of 86%**, showing balanced performance across both classes.
- **Class 0 (True News)** shows slightly better precision and F1-score than class 1.
- The macro and weighted averages are consistent, indicating **no significant class imbalance impact**.
- Performance is comparable across both real and fake news, suggesting that the tree structure captures meaningful patterns in the data.

Confusion Matrix– Random Forest Model

Class	Precision	Recall	F1-Score	Support
0 (True News)	0.94	0.93	0.94	7103
1 (Fake News)	0.92	0.93	0.93	6367
Accuracy			0.93	13470
Macro Avg	0.93	0.93	0.93	13470
Weighted Avg	0.93	0.93	0.93	13470

- The model achieves a **high accuracy of 93%**, outperforming both Logistic Regression and Decision Tree models.
- **Balanced precision and recall** for both real and fake news indicates reliable classification.
- Macro and weighted averages match closely, showing **consistent performance across classes**.

Insights

Model	Accuracy	Precision (Class 0 / 1)	Recall (Class 0 / 1)	F1-Score (Class 0 / 1)
Logistic Regression	0.84	0.90 / 0.78	0.82 / 0.88	0.86 / 0.83
Decision Tree	0.86	0.88 / 0.85	0.86 / 0.86	0.87 / 0.85
[Best Model – e.g., Random Forest]	0.93	0.94 / 0.92	0.93 / 0.93	0.94 / 0.93

- **Logistic Regression** performs well but has lower precision on fake news (Class 1), indicating more false positives.
- **Decision Tree** improves both class-wise precision and recall over logistic regression, showing better balance.
- **Best Model (Random Forest)** achieves the **highest accuracy and class balance**

Outcomes

- Using **Word2Vec embeddings** added meaningful context to the text by capturing the relationships between words, which clearly helped improve classification performance.
- The **Random Forest model** delivered the **best performance**, achieving a high accuracy of **93%** with well-balanced precision and recall for both true and fake news.
- When compared to **Logistic Regression (84%)** and the **Decision Tree (86%)**, Random Forest showed stronger generalization and better reliability in identifying fake content.