# Fake News Detection- Report

(By Priyanka Gulati)

## 1. Executive Summary

This report outlines the approach and findings of a fake news detection system using semantic classification techniques. By leveraging the Word2Vec model and traditional supervised learning algorithms like Logistic Regression, Decision Tree, and Random Forest, the system aims to classify news articles as either true or fake based on their textual content. The analysis is supported by extensive data preprocessing, exploratory data analysis, and model evaluation to ensure reliability and performance.

Analysis shows that true news articles tend to use formal and institutional language, while fake news often relies on emotionally charged, dramatic, or sensational wording. Logistic Regression delivered a strong performance, with high accuracy and recall, making it ideal for identifying true news articles. The semantic classification approach, powered by Word2Vec embeddings and robust preprocessing, effectively captures linguistic patterns to distinguish fake from true news.

## 2. Business Objective

The growing dissemination of fake news poses a threat to public trust and informed decision-making. This project addresses the challenge by building a semantic classification model to distinguish between fake and true news articles. The primary goals are:

- To identify recurring linguistic and semantic patterns in news texts.
- To accurately classify news using supervised learning techniques.
- To compare different models and choose the most effective one based on relevant evaluation metrics.

## 3. Data Preparation

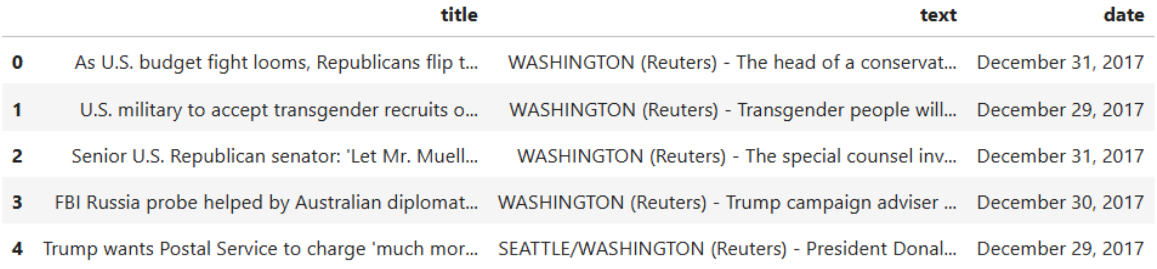
### 3.1 Data Understanding

Dataset is analyzed to understand the basic structure and key statistics. The data has 2 files, containing True News and Fake News respectively. Primary details are as follows:
Datasets Used:

- True.csv: 21,417 true news article rows
- Fake.csv: 23,523(23,502 populated) fake news article rows

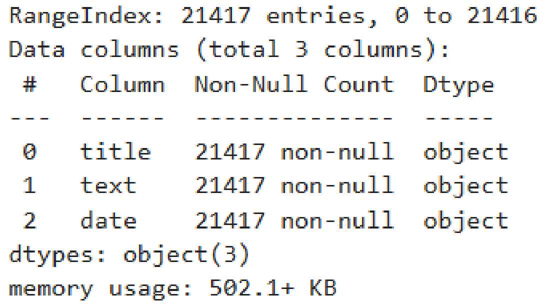
Key Attributes:

- Title: News headline.
- Text: Main content of the article.
- Date: Publication date.

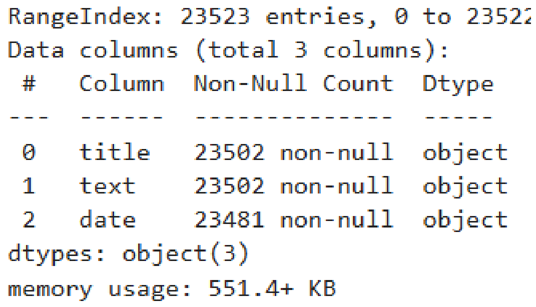


| | title | text | date |
|---|---|---|---|
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | December 31, 2017 |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | December 29, 2017 |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | December 31, 2017 |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | December 30, 2017 |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | December 29, 2017 |

True News articles have no null values whereas fake news articles have null values. These will be properly identified and treated. (Details in section 3.4)



```
RangeIndex: 21417 entries, 0 to 21416
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   title   21417 non-null  object
 1   text    21417 non-null  object
 2   date    21417 non-null  object
dtypes: object(3)
memory usage: 502.1+ KB
```



```
RangeIndex: 23523 entries, 0 to 23522
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   title   23502 non-null  object
 1   text    23502 non-null  object
 2   date    23481 non-null  object
dtypes: object(3)
memory usage: 551.4+ KB
```

The data also had duplicate values. 217 observations in true news and 5601 observations in fake news are duplicates. Duplicates are dropped as they do not add any additional information to our analysis.

Key Attributes are further analysed to identify the unique values in each column.

Unique values in true news:

```
title    20826
text     21192
date       716
```

Unique values in fake news:

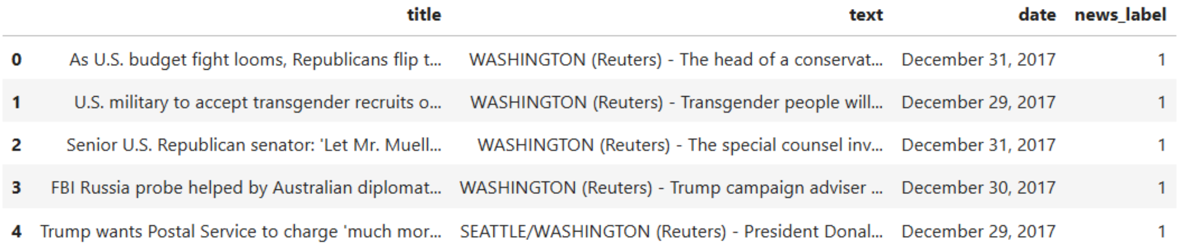
```
title    17914
text     17466
date      1692
```

## 3.2 Add New Column

Appropriate label columns are added in true news and fake news, marking true news as 1 and fake news as 0.

## 3.3 Merge Data Frames

True News and fake news data frames are combined into a single data frame using pd.concat.



| | title | text | date | news_label |
|---|---|---|---|---|
| 0 | As U.S. budget fight looms, Republicans flip t... | WASHINGTON (Reuters) - The head of a conservat... | December 31, 2017 | 1 |
| 1 | U.S. military to accept transgender recruits o... | WASHINGTON (Reuters) - Transgender people will... | December 29, 2017 | 1 |
| 2 | Senior U.S. Republican senator: 'Let Mr. Muell... | WASHINGTON (Reuters) - The special counsel inv... | December 31, 2017 | 1 |
| 3 | FBI Russia probe helped by Australian diplomat... | WASHINGTON (Reuters) - Trump campaign adviser ... | December 30, 2017 | 1 |
| 4 | Trump wants Postal Service to charge 'much mor... | SEATTLE/WASHINGTON (Reuters) - President Donal... | December 29, 2017 | 1 |

Updated dataset has the additional label column and has 39,122 rows (17922+21200 rows post removal of duplicates).

## 3.4 Handling Null values

Null values were observed in the merged data frame. The count and the percentage of null values are as follows:

Count of missing values in each column:

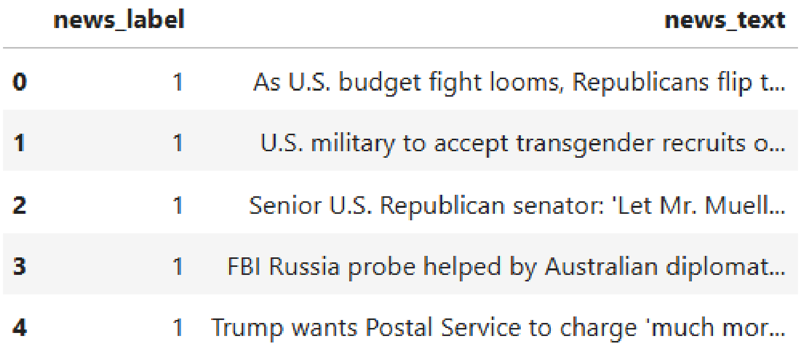
```
title         1
text          1
date         12
news_label    0
```

```
Percentage of missing values in each column:
title        0.003
text          0.003
date          0.031
news_label    0.000
```

As the number of null values are insignificant, null value rows are dropped from the data. Post the null value treatment there are 39,110 rows and 4 columns in the data.

### 3.5 Merge the relevant columns and drop the rest from the data frame

Text columns (Title and Text) are combined into a new column news_text. Different text fields are combined into one cohesive text input for semantic analysis as word2vec models work better on continuous text rather than fragments. Irrelevant columns are removed from the data frame.



| | news_label | news_text |
|---|---|---|
| 0 | 1 | As U.S. budget fight looms, Republicans flip t... |
| 1 | 1 | U.S. military to accept transgender recruits o... |
| 2 | 1 | Senior U.S. Republican senator: 'Let Mr. Muell... |
| 3 | 1 | FBI Russia probe helped by Australian diplomat... |
| 4 | 1 | Trump wants Postal Service to charge 'much mor... |

## 4. Text Preprocessing

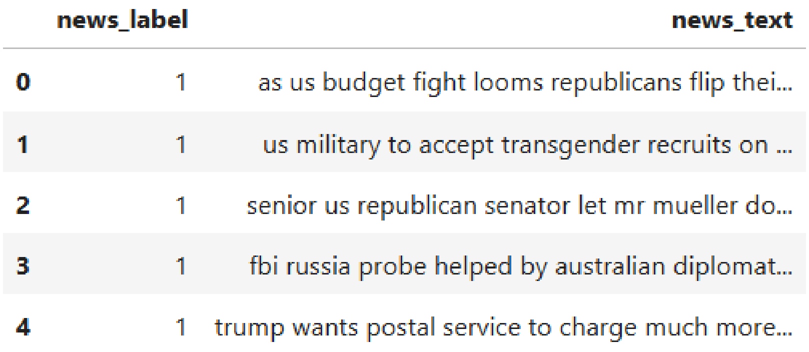
News Text is cleaned by performing the below steps:

1. Making the text lowercase
2. Removing text in square brackets
3. Removing punctuation
4. Removing words containing numbers

After cleaning the text data, POS tagging and lemmatization is done on the cleaned news text, and all words that are not tagged as NN or NNS are removed.

### 4.1 Text Cleaning

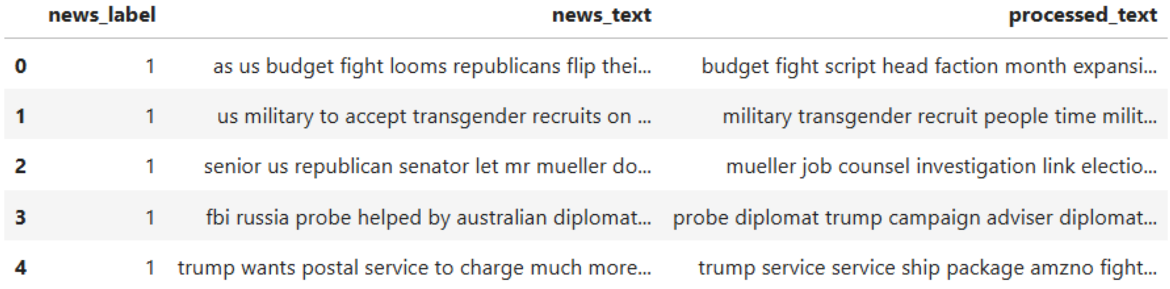
Clean text function is written where all the text cleaning steps (lowercasing, removing square brackets, removing punctuation, removing words with numbers) are performed. The text is cleaned using the function and a new data frame with clean text is created.



| | news_label | news_text |
|---|---|---|
| **0** | 1 | as us budget fight looms republicans flip thei... |
| **1** | 1 | us military to accept transgender recruits on ... |
| **2** | 1 | senior us republican senator let mr mueller do... |
| **3** | 1 | fbi russia probe helped by australian diplomat... |
| **4** | 1 | trump wants postal service to charge much more... |

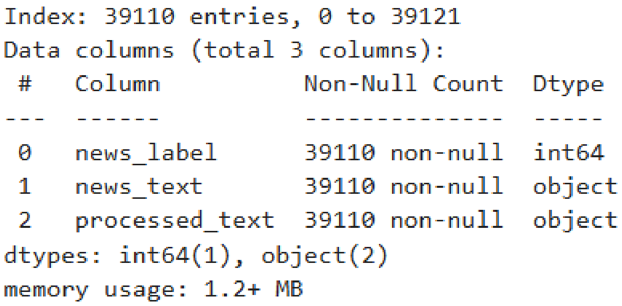
### 4.2 POS Tagging & Lemmatization

Function is created to select the lemmatized version of words post removal of stop words and filtering for noun POS tags (NN and NNS). Additional column (processed_text) is created in the data containing the lemmatized (POS and stop words filtered) text.



| | news_label | news_text | processed_text |
|---|---|---|---|
| **0** | 1 | as us budget fight looms republicans flip thei... | budget fight script head faction month expansi... |
| **1** | 1 | us military to accept transgender recruits on ... | military transgender recruit people time milit... |
| **2** | 1 | senior us republican senator let mr mueller do... | mueller job counsel investigation link electio... |
| **3** | 1 | fbi russia probe helped by australian diplomat... | probe diplomat trump campaign adviser diplomat... |
| **4** | 1 | trump wants postal service to charge much more... | trump service service ship package amzno fight... |

Now, there is clean dataset which has appropriate datatypes and no null values.



```
Index: 39110 entries, 0 to 39121
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   news_label      39110 non-null  int64 
 1   news_text       39110 non-null  object
 2   processed_text  39110 non-null  object
dtypes: int64(1), object(2)
memory usage: 1.2+ MB
```

## 5. <u>Train Validation Split</u>

Data is split into train and validation data with train data having 70 % of the observations and validation data having 30 % of the observations. Post the split, train data has 27,377 rows and validation data has 11,733 rows.

## 6. <u>Exploratory Data Analysis on Training Data</u>

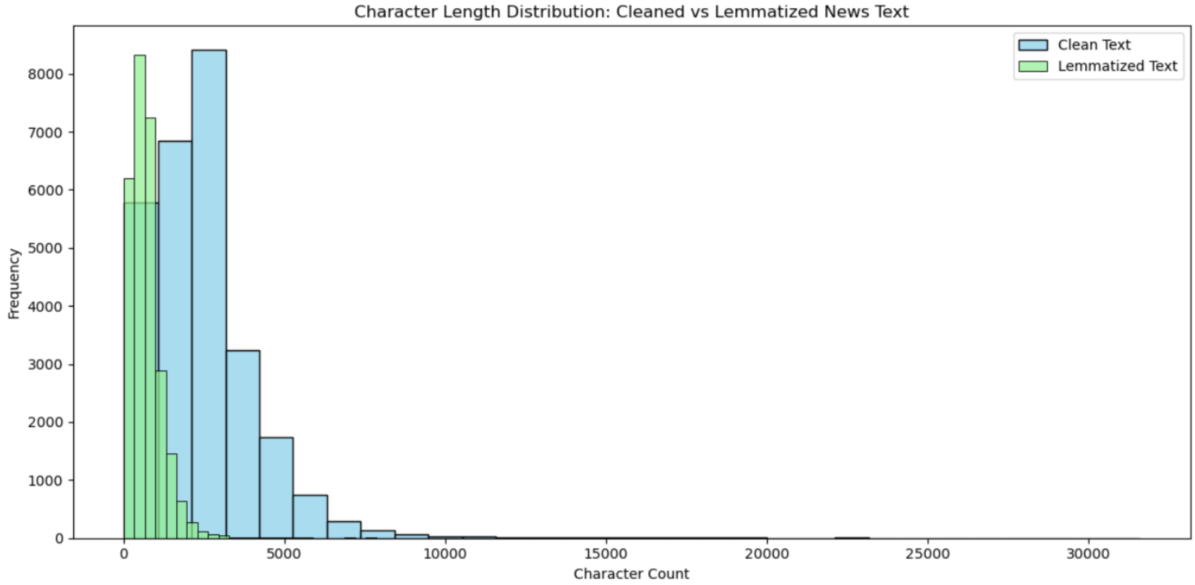
EDA is performed on cleaned and pre-processed texts to get familiar with the training data by performing the below steps:

- Visualising the training data according to the character length of cleaned news text and lemmatized news text with POS tags removed
- Using a word cloud, finding the top 40 words by frequency in true and fake news separately

- Finding the top unigrams, bigrams and trigrams by frequency in true and fake news separately

### 6.1 Visualise character lengths of cleaned news text and lemmatized news text with POS tags removed

Character lengths are calculated for the cleaned news text and lemmatized news text. The distributions are plotted on the same graph for comparison and overlaps and peak differences are observed to understand text preprocessing's impact on text length.



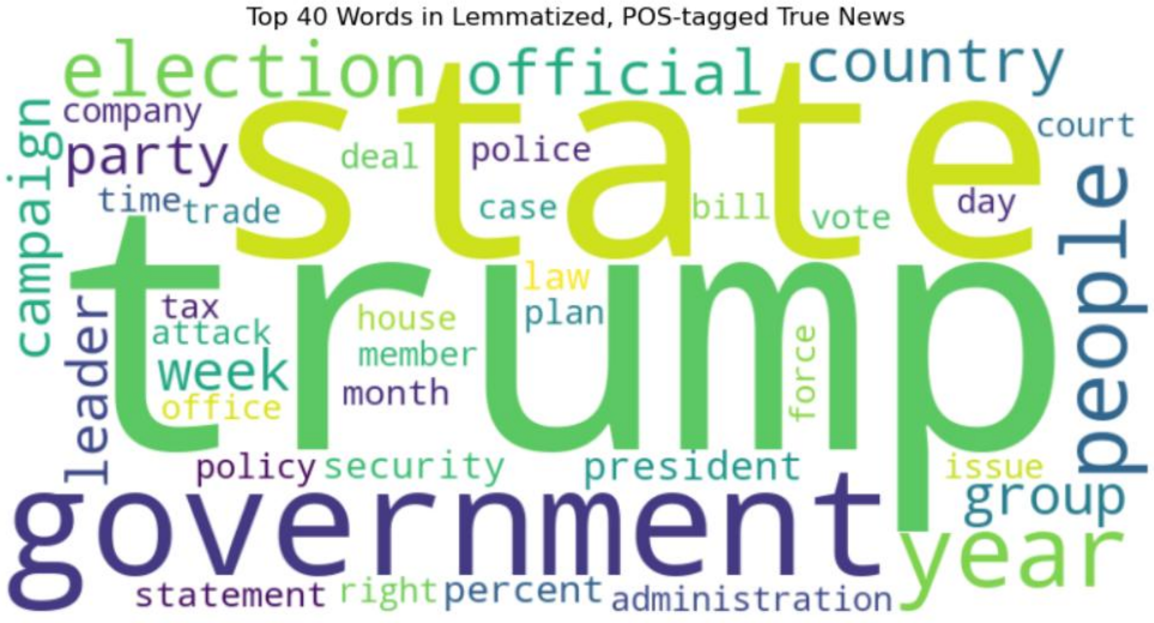
**Observations**

- The cleaned text (blue) has a wider distribution, stretching much farther on the right (some texts >20,000 characters).
- The lemmatized POS-tagged text (green) is tightly concentrated toward the left - meaning text is significantly shorter after filtering for nouns, removing stop words and lemmatizing the text.
- The peak for lemmatized version is fairly below the peak for cleaned text.
- There's overlap, but the lemmatized version compresses the text a lot more, as expected.

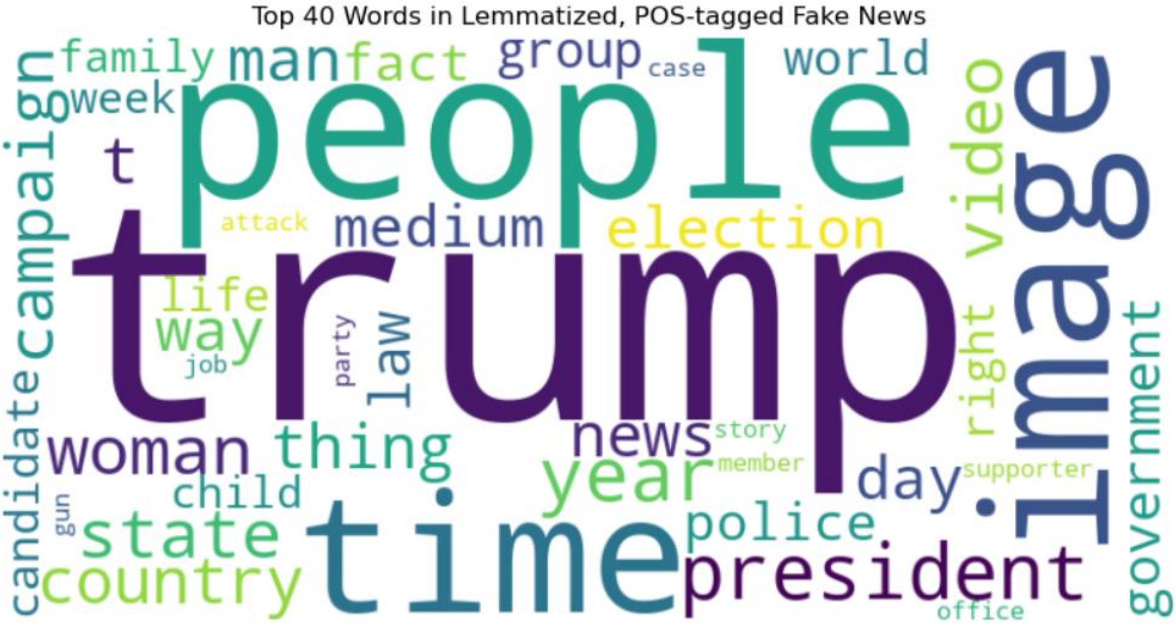
### 6.2 Find and display the top 40 words by frequency among true and fake news in Training data after processing the text

True news and Fake news are identified using the labels 1 and 0. The text is converted to strings and joined to create a common corpus of all news articles for true news and false news respectively. Corpus is split into words and the frequency is calculated to identify the top 40 words having highest frequency. Word clouds are created for top 40 frequent words for true and fake news.

Word Cloud for True News



Word Cloud for Fake News



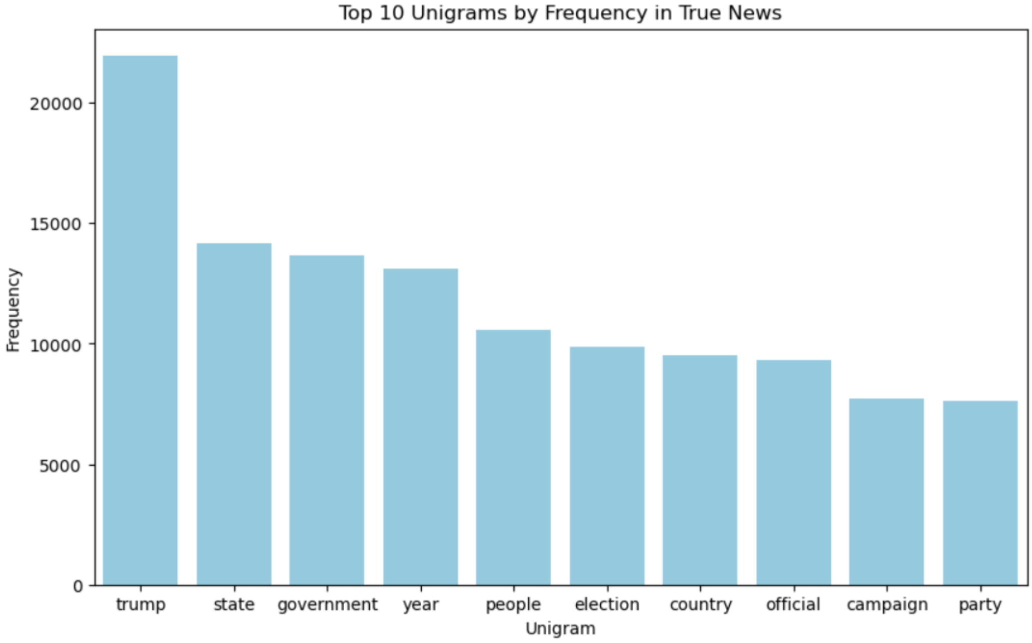
**Observations**

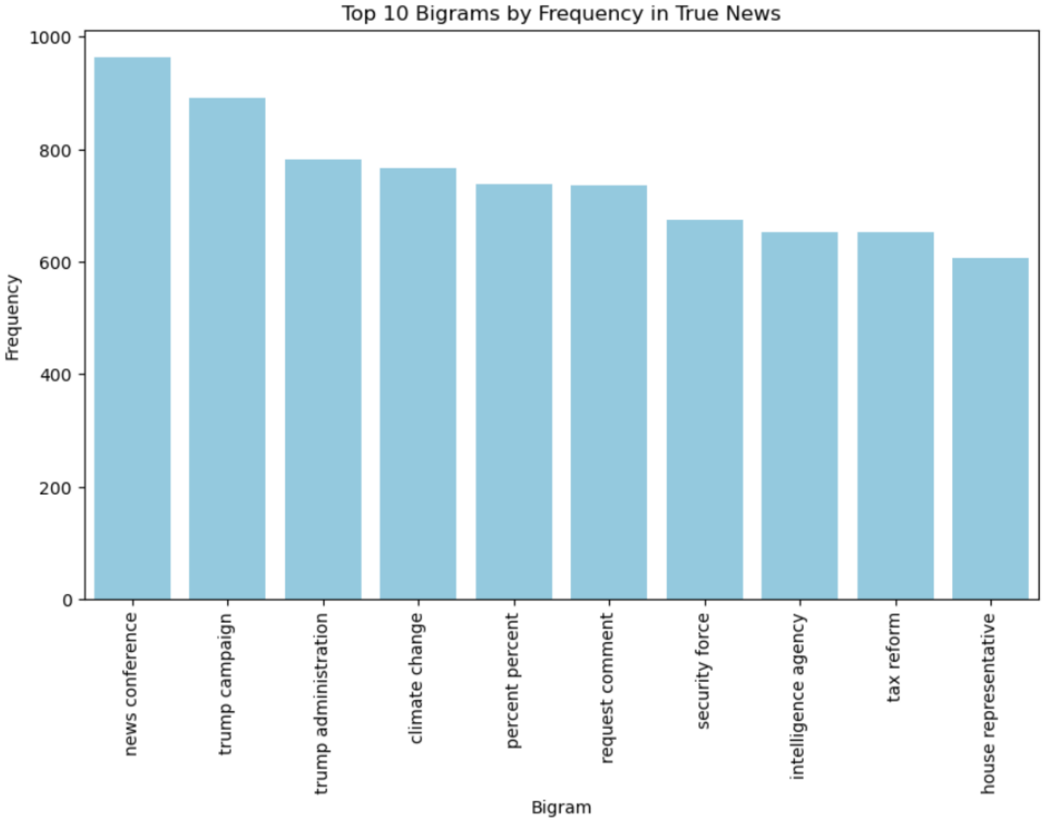
- Both True and Fake News mention Trump, State, Government, Election, Year and People. These are likely general political topics covered in both kinds of news.
- Fake News focuses more on emotionally charged words like attack, image, gun, woman, child. This could be an effort to sensationalise and create emotionally provocative headlines.
- True News contains more formal and institutional language with usage of words like official, administration, statement, policy, court.

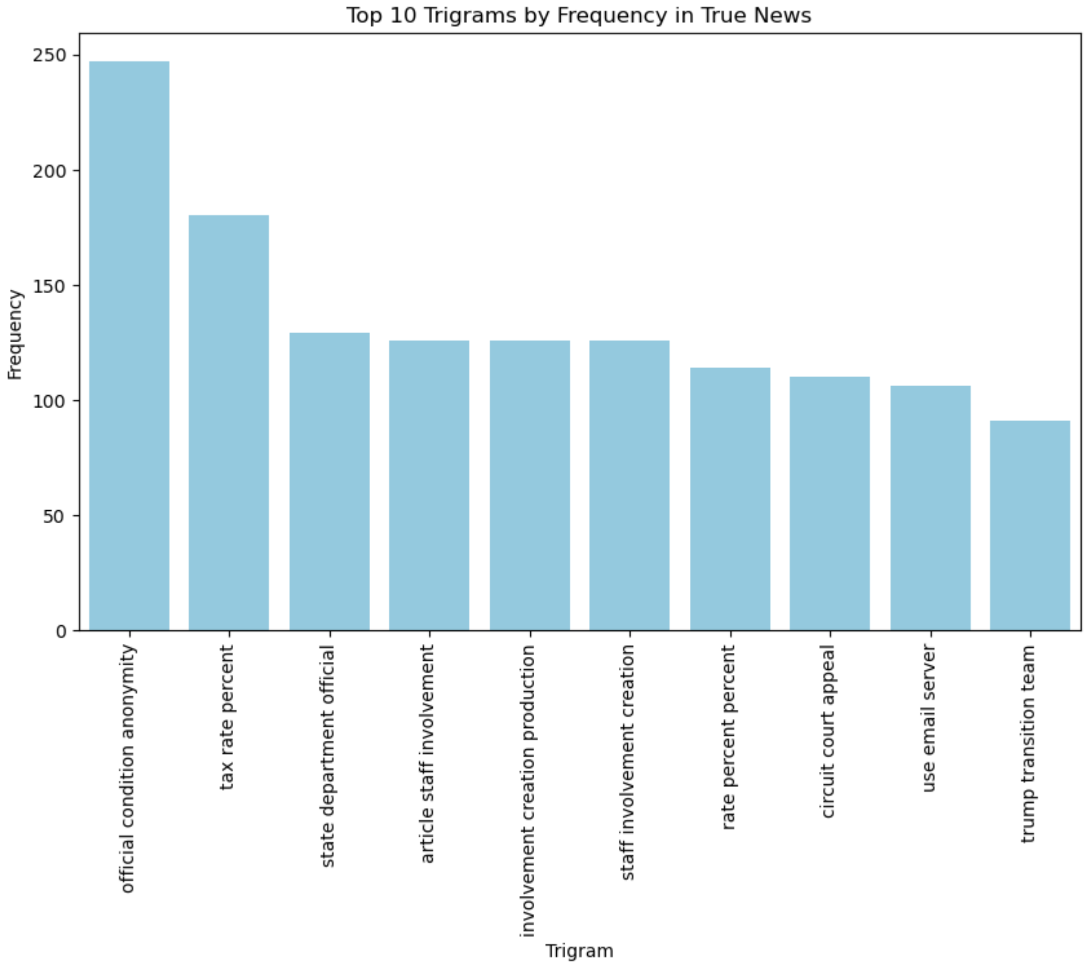
### 6.3 Find and display the top unigrams, bigrams and trigrams by frequency in true news and fake news after processing the text

Function is created to identify the top n-grams using Count Vectorizer. The data is verified that it contains no null values. Top 10 unigrams, bigrams and trigrams are identified and plotted on a bar chart for true news and fake news.

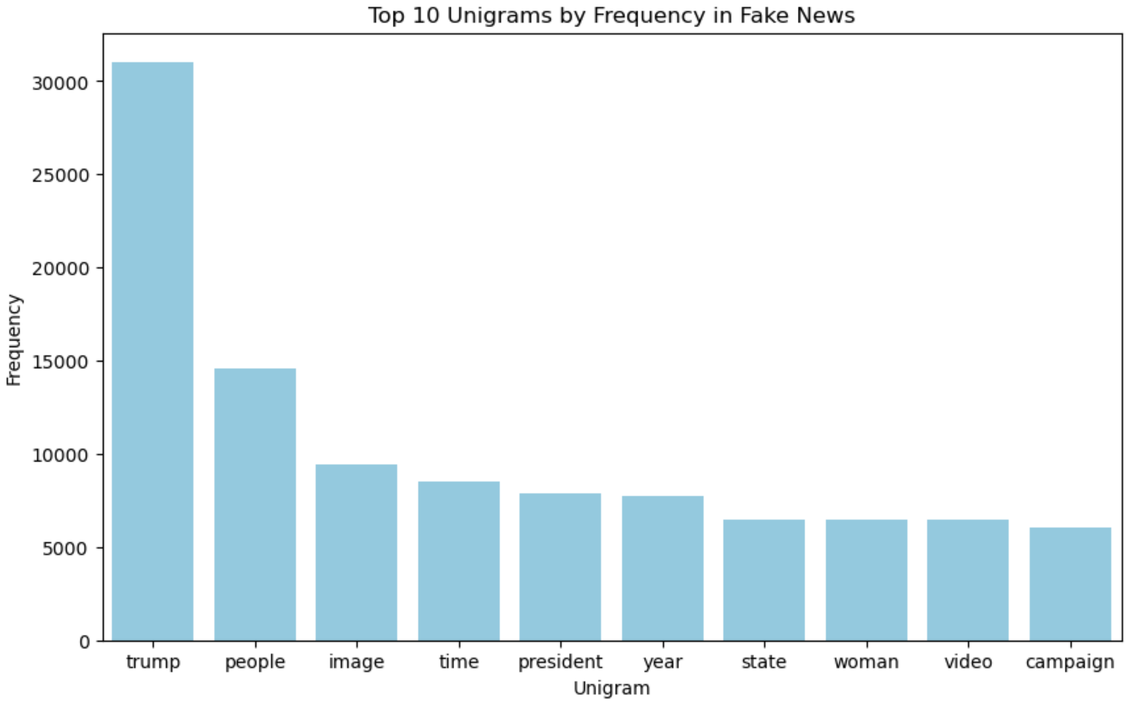
Top 10 Unigram, Bigram and Trigram for True News

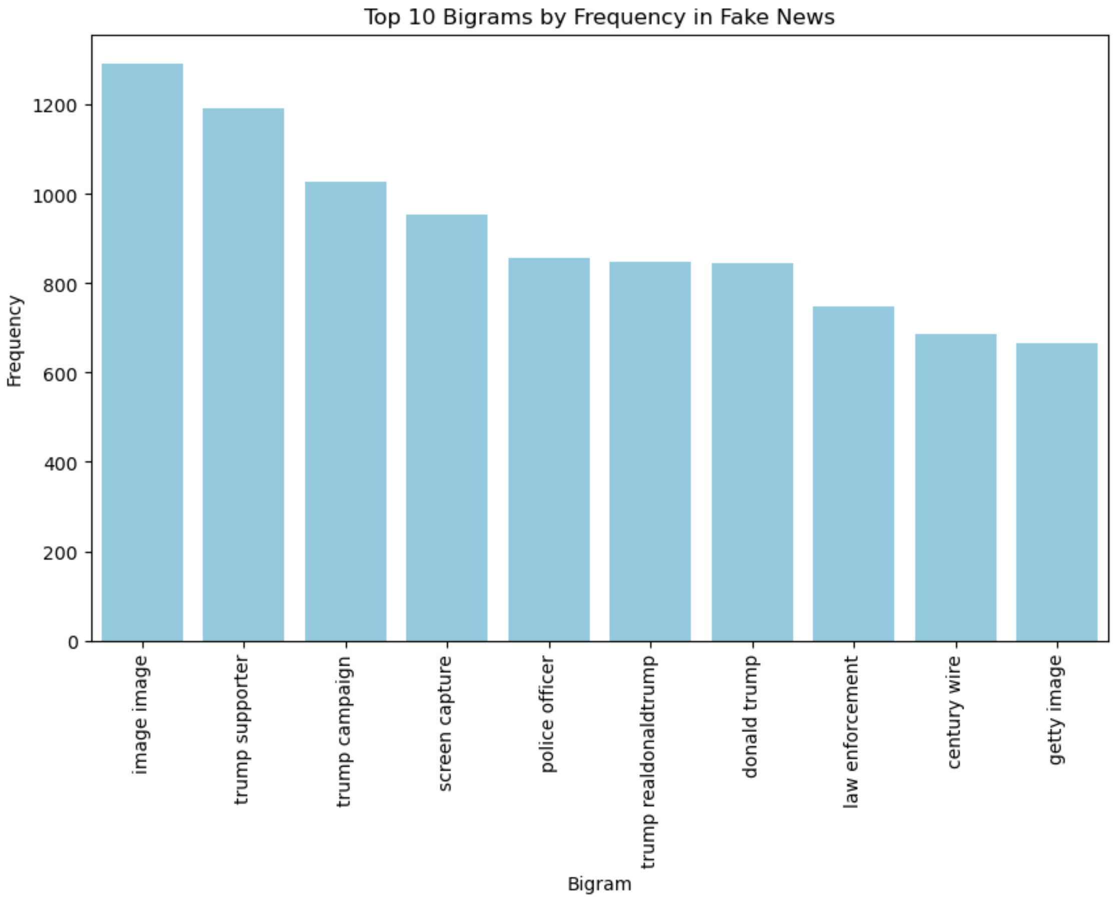


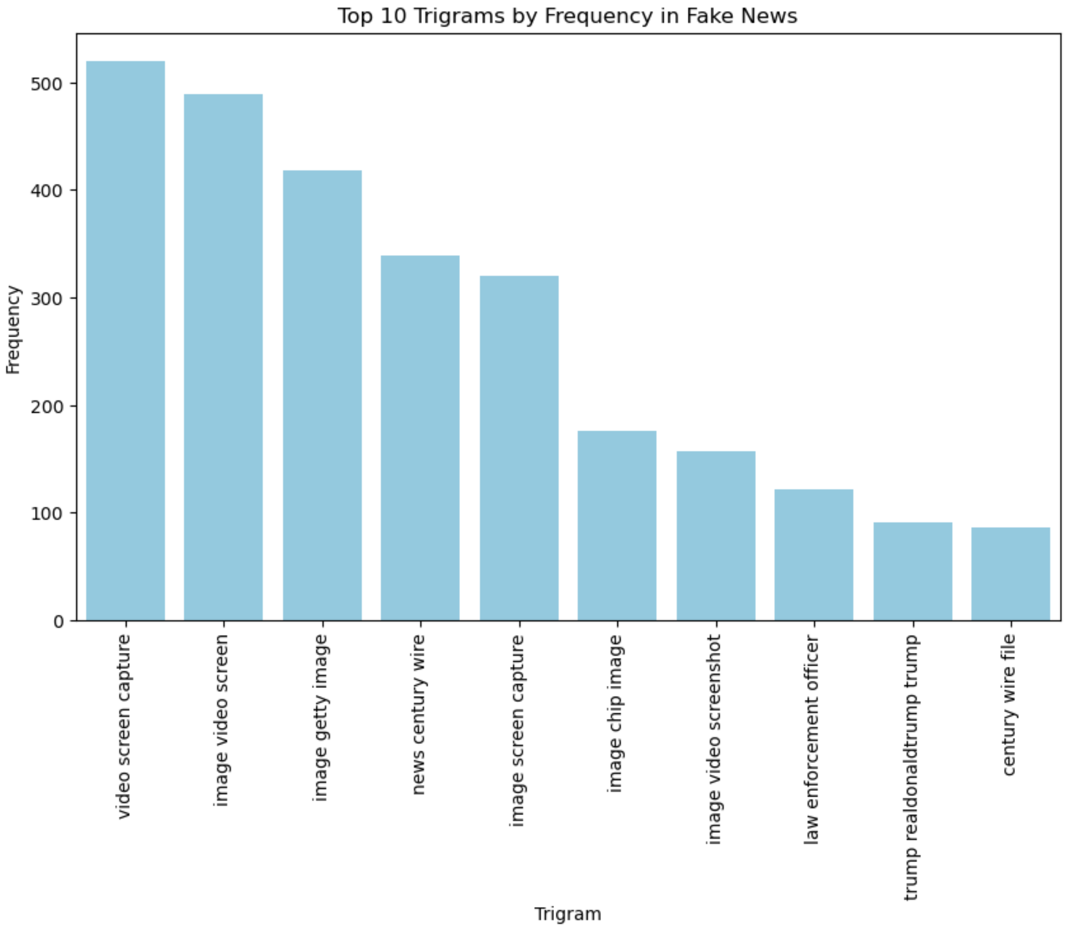




## Top 10 Unigram, Bigram and Trigram for Fake News




Top 10 Bigrams by Frequency in Fake News
Frequency
0
200
400
600
800
1000
1200
image image
trump supporter
trump campaign
screen capture
police officer
trump realdonaldtrump
donald trump
law enforcement
century wire
getty image
Bigram


Top 10 Trigrams by Frequency in Fake News
Frequency
0
100
200
300
400
500
video screen capture
image video screen
image getty image
news century wire
image screen capture
image chip image
image video screenshot
law enforcement officer
trump realdonaldtrump trump
century wire file
Trigram

**Observations**

- Unigrams- True News uses formal terms like government, election etc. Fake News leans towards visual and sensational words like Image, Video etc. There are some common themes like Trump, State indicating the nature of political data.
- Bigrams-True News bigrams are policy/government related (like trump administration, climate change). Fake News features are repetitive or viral terms (like image image, trump supporter)
- Trigrams- True News has structured phrases (like state department official). Fake news trigrams are primarily media related (like video screen capture, image getty image), hinting visual or fabricated content.

## 7. Exploratory Data Analysis on Validation Data (Optional)

Similar EDA is performed on the validation data (like Training Data) by creating distributions, word clouds and n-grams. In order to keep the report concise, listing only the primary observation from the EDA of validation data.

**Distribution of Cleaned Text and Lemmatized Text**

Validation Data shows similar pattern in character length distributions (like Train Data), where clean text has a much wider distribution than the lemmatized text. Peak for lemmatized text is fairly below the cleaned text.

**Word cloud**

- Even in Validation Data, Fake News focuses more on emotionally charged words like attack, image, video, gun, woman, child. This could be an effort to sensationalise and create emotionally provocative headlines.
- Even in Validation Data, True News contains more formal and institutional language with usage of words like administration, law, statement, policy, court.

**N-Grams**

- Unigrams- Even in Validation data, True News uses formal terms like government, election etc. Fake News leans towards visual and sensational words like Image, Video etc. There are some common themes like Trump, State indicating the nature of political data.
- Bigrams-Even in Validation data, True News bigrams are policy/government related and formal (like trump administration, news conference). Fake News features are repetitive or viral terms (like image image, trump supporter)
- Trigrams- Even in Validation data, True News has structured phrases (like state department official). Fake news trigrams are primarily media related (like video screen capture, image getty image), hinting visual or fabricated content.

## 8. Feature Extraction

### 8.1 Initialize Word2Vec Model

Word2Vec model is initialized by to loading "word2vec-google-news-300" which is a pretrained word embedding model developed by Google, trained on 100 billion words from Google News.

### 8.2 Extract vectors for cleaned news data

Function is created to split each news article in the input text data to tokens and convert each word into a vector using a pretrained Word2Vec model. For each article, it averages the vectors of the known words to create a single vector representation. If none of the words in an article are found in the model's vocabulary, a zero vector is assigned instead.

Lemmatized text is converted to vector form for training and validation data and saved as X_train and X_val respectively. Target Variable (news_label) for training and validation data is saved as y_train and y_val respectively.

## 9. Model Training & Evaluation

Supervised Models are created to classify the news as fake or true.

### 9.1 Import models and evaluation metrics

Libraries are loaded to run the Logistic Regression, Decision Tree and Random Forest Classifier models and create classification reports and evaluation metrics like accuracy, precision, recall and f1 score.

### 9.2 Build Logistic Regression Model

Logistic Regression model is initialised, tuned for hyper-parameters, fitted on train data and predictions are generated for validation data. Accuracy, Precision, Recall & F1 score is calculated. Classification Report is also generated.

Accuracy: 0.9095
Precision: 0.9114
Recall: 0.9223
F1 Score: 0.9168

- **90.95%** of the total news articles are correctly classified (Accuracy).
- **91.14%** of the news articles predicted as true are actually true (Precision).
- **92.23%** of the actual true news articles are correctly predicted as true (Recall).
- **91.68%** is the F1 Score, which is the harmonic mean of precision and recall — indicating a good balance between both.

Classification Report:

```
         precision    recall  f1-score   support

       0       0.91      0.89      0.90      5386
       1       0.91      0.92      0.92      6347

accuracy                           0.91     11733
   macro avg       0.91      0.91      0.91     11733
weighted avg       0.91      0.91      0.91     11733
```

- The model achieved an overall accuracy of 91%.
- True News was classified slightly better, with 92% F1-score, compared to 90% for Class 0.
- Both macro and weighted averages are balanced at 91%, indicating consistent performance across both classes.

### 9.3 Build Decision Tree Model

Decision Tree model is initialised, tuned for hyper-parameters, fitted on train data and predictions are generated for validation data. Accuracy, Precision, Recall & F1 score is calculated. Classification Report is also generated.

Accuracy: 0.8183
Precision: 0.8307
Recall: 0.8341
F1 Score: 0.8324

- **81.83%** of the total news articles are correctly classified (Accuracy).
- **83.07%** of the news articles predicted as true are actually true (Precision).
- **83.41%** of the actual true news articles are correctly predicted as true (Recall).
- **83.24%** is the F1 Score, which is the harmonic mean of precision and recall — indicating a good balance between both.

```
Classification Report:
               precision    recall  f1-score   support

           0       0.80      0.80      0.80      5386
           1       0.83      0.83      0.83      6347

    accuracy                           0.82     11733
   macro avg       0.82      0.82      0.82     11733
weighted avg       0.82      0.82      0.82     11733
```

- The model achieved an overall accuracy of 82%.
- True News was classified slightly better, with 83% F1-score, compared to 80% for Class 0.
- Both macro and weighted averages are balanced at 82%, indicating consistent performance across both classes.

### 9.4 Build Random Forest Model

Random Forest model is initialised, tuned for hyper-parameters, fitted on train data and predictions are generated for validation data. Accuracy, Precision, Recall & F1 score is calculated. Classification Report is also generated.

Accuracy: 0.878
Precision: 0.8717
Recall: 0.908
F1 Score: 0.8895

- **87.8%** of the total news articles are correctly classified (Accuracy).
- **87.17%** of the news articles predicted as true are actually true (Precision).
- **90.8%** of the actual true news articles are correctly predicted as true (Recall).
- **88.95%** is the F1 Score, which is the harmonic mean of precision and recall — indicating a good balance between both.

```
Classification Report:
               precision    recall  f1-score   support
```

```
           0       0.89      0.84      0.86      5386
           1       0.87      0.91      0.89      6347

    accuracy                           0.88     11733
   macro avg       0.88      0.88      0.88     11733
weighted avg       0.88      0.88      0.88     11733
```

- The model achieved an overall accuracy of 88%.
- True News was classified slightly better, with 89% F1-score, compared to 86% for Class 0.
- Both macro and weighted averages are balanced at 88%, indicating consistent performance across both classes.

**Model Selection**

- Logistic Regression performed the best overall, with the highest F1-score (91.68%).
- Random Forest performed well too, especially in Recall (90.8%), though it lagged slightly behind Logistic Regression in Precision and F1-score.
- Decision Tree was relatively weaker in performance but was still able to provide reasonable classifications.

## 10.Conclusion

**Key Observations:**

- True news articles tend to use formal, institutional language, while fake news often relies on emotionally charged, dramatic, or sensational wording.
- N-gram analysis highlights that true news focuses on structured governance and policy related words, whereas fake news frequently includes visual and viral terms like "image" and "video."

**Best Model:**

- Logistic Regression delivered the strongest performance, with high evaluation metrics making it useful for identification of true news articles.
- It not only has balanced evaluation metrics but also is interpretable, allowing insights into which features influence predictions - a valuable trait in domains requiring transparency.

**Impact:**

- The semantic classification approach, using Word2Vec embeddings and robust preprocessing, effectively captures linguistic patterns to distinguish fake from true news.
- The solution is scalable and adaptable to other misinformation domains and suitable for integration into real-time platforms - helping to enhance public trust and reduce misinformation.