Fake News Detection Analysis

November 2025

Prepared by: Suchindran Kannan, Roma Agrawal

Business Objective

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

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, we developed 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 performed:

We performed the following tasks to complete the assignment:

1. Data Preparation
2. Text Preprocessing
3. Train Validation Split
4. EDA on Training Data
5. EDA on Validation Data [Optional]
6. Feature Extraction
7. Model Training and Evaluation

Data Dictionary

For this assignment, you woredk with 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.

Let’s begin:

After installing and importing all the necessary libraries, we loaded the True.csv and Fake.csv files as DataFrames.

1. Data Preparation

Inspected the DataFrame with True News and Fake News to understand the given data.

Add new column

added new column `news_label` to both the DataFrames and assign labels

[10…]
```
# Add a new column 'news_label' to the true news DataFrame and assign the label "1" to indicate that t

true_df['news_label'] = 1

# Add a new column 'news_label' to the fake news DataFrame and assign the label "0" to indicate that t

fake_df['news_label'] = 0

# Verify the changes
true_df.head(), fake_df.head()
```

[10…]
```
(                                                title  \
 0       As U.S. budget fight looms, Republicans flip their fiscal script
 1       U.S. military to accept transgender recruits on Monday: Pentagon
 2           Senior U.S. Republican senator: 'Let Mr. Mueller do his job'
 3            FBI Russia probe helped by Australian diplomat tip-off: NYT
 4  Trump wants Postal Service to charge 'much more' for Amazon shipments


text  \
 0  WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who vot
ed this month for a huge expansion of the national debt to pay for tax cuts, called himself a “fis...
 1  WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S.
military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after Pres...
 2  WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Tru
mp’s 2016 election campaign should continue without interference in 2018, despite calls from some ...
 3  WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in M
ay 2016 that Russia had political dirt on Democratic presidential candidate Hillary Clinton, the N...
 4  SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday
to charge “much more” to ship packages for Amazon (AMZN.O), picking another fight with an online r...

                 date  news_label
 0  December 31, 2017            1
 1  December 29, 2017            1
 2  December 31, 2017            1
 3  December 30, 2017            1
 4  December 29, 2017            1  ,
                                                                                title  \
 0              Donald Trump Sends Out Embarrassing New Year’s Eve Message; This is Disturbing
 1                        Drunk Bragging Trump Staffer Started Russian Collusion Investigation
 2   Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People ‘In The Eye’
 3               Trump Is So Obsessed He Even Has Obama’s Name Coded Into His Website (IMAGES)
 4                       Pope Francis Just Called Out Donald Trump During His Christmas Speech


text  \
 0  Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Instead, he h
ad to give a shout out to his enemies, haters and  the very dishonest fake news media.  The former...
 1  House Intelligence Committee Chairman Devin Nunes is going to have a bad day. He s been under the
assumption, like many of us, that the Christopher Steele-dossier was what prompted the Russia inve...
 2  On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered fo
r Homeland Security Secretary in Donald Trump s administration, has an email scandal of his own.In...
 3  On Christmas day, Donald Trump announced that he would  be back to work  the following day, but he
is golfing for the fourth day in a row. The former reality show star blasted former President Bar...
 4  Pope Francis used his annual Christmas Day message to rebuke Donald Trump without even mentioning
his name. The Pope delivered his message just days after members of the United Nations condemned T...

                 date  news_label
 0  December 31, 2017            0
 1  December 31, 2017            0
 2  December 30, 2017            0
 3  December 29, 2017            0
 4  December 25, 2017            0  )
```

1.2 Merge DataFrames

Create a new Dataframe by merging True and Fake DataFrames

[11…]
```
# Combine the true and fake news DataFrames into a single DataFrame
news_df = pd.concat([true_df, fake_df], axis=0, ignore_index=True)
```

[12…]
```
# Display the first 5 rows of the combined DataFrame to verify the result
news_df.head()
```

[12…]

	title	text	date	news_label
0	As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a “fis...	December 31, 2017	1
1	U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as ordered by federal courts, the Pentagon said on Friday, after Pres...	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Trump’s 2016 election campaign should continue without interference in 2018, despite calls from some ...	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat tip-off: NYT	WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt on Democratic presidential candidate Hillary Clinton, the N...	December 30, 2017	1
4	Trump wants Postal Service to charge 'much more' for Amazon shipments	SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday to charge “much more” to ship packages for Amazon (AMZN.O), picking another fight with an online r...	December 29, 2017	1

1.3 Handle the null values

Check for null values and handle it by imputation or dropping the null values

[13…]
```
# Check Presence of Null Values
news_df.isnull().sum()
```

[13…]

	0
title	21
text	21
date	42
news_label	0

dtype: int64

[14…]
```
# Handle Rows with Null Values

news_df = news_df.dropna()

# Verify again
news_df.isnull().sum()
```

[14…]

	0
title	0
text	0
date	0
news_label	0

dtype: int64

1.4 Merge the relevant columns and drop the rest from the DataFrame

Combine the relevant columns into a new column `news_text` and then drop irrelevant columns from the DataFrame

[15…]
```
# Combine the relevant columns into a new column 'news_text' by joining their values with a space

news_df['news_text'] = news_df['title'].astype(str) + " " + news_df['text'].astype(str) + " " + news_d

# Drop the irrelevant columns from the DataFrame as they are no longer needed

news_df = news_df.drop(columns=['title', 'text', 'date'])

# Display the first 5 rows of the updated DataFrame to check the result

news_df.head()
```

[15…]

	news_label	news_text
0	1	As U.S. budget fight looms, Republicans flip their fiscal script WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansio...
1	1	U.S. military to accept transgender recruits on Monday: Pentagon WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as o...
2	1	Senior U.S. Republican senator: 'Let Mr. Mueller do his job' WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Trump's 2016 election campaign should co...
3	1	FBI Russia probe helped by Australian diplomat tip-off: NYT WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt...
4	1	Trump wants Postal Service to charge 'much more' for Amazon shipments SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday to charge "much more" to shi...

2. Text Preprocessing

2.1 Text Cleaning

Created a new DataFrame to store the processed data and wrote the function to clean the text and remove all the unnecessary elements. Later we applied a function to clean the news text and store the cleaned text in a new column within the new DataFrame.

Here is how latest results looks like:

[18]:

	raw_text	clean_text
0	As U.S. budget fight looms, Republicans flip their fiscal script WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansio...	as us budget fight looms republicans flip their fiscal script washington reuters the head of a conservative republican faction in the us congress who voted this month for a huge expansion of the ...
1	U.S. military to accept transgender recruits on Monday: Pentagon WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as o...	us military to accept transgender recruits on monday pentagon washington reuters transgender people will be allowed for the first time to enlist in the us military starting on monday as ordered b...
2	Senior U.S. Republican senator: 'Let Mr. Mueller do his job' WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Trump's 2016 election campaign should co...	senior us republican senator let mr mueller do his job washington reuters the special counsel investigation of links between russia and president trump's election campaign should continue withou...
3	FBI Russia probe helped by Australian diplomat tip-off: NYT WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt...	fbi russia probe helped by australian diplomat tipoff nyt washington reuters trump campaign adviser george papadopoulos told an australian diplomat in may that russia had political dirt on democ...
4	Trump wants Postal Service to charge 'much more' for Amazon shipments SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday to charge "much more" to shi...	trump wants postal service to charge much more for amazon shipments seattlewashington reuters president donald trump called on the us postal service on friday to charge "much more" to ship packag...

2.2 POS Tagging and Lemmatization

Wrote the function for POS tagging and lemmatization, filtering stopwords and keeping only NN and NNS tags. Applied the POS tagging and lemmatization function to cleaned text and store it in a new column within the new DataFrame And Saved the Cleaned data as a csv file.

Here is how latest results looks like:

[25]:
```
# Check the first few rows of the DataFrame
df_clean.head()
```

[25]:

	raw_text	news_label	clean_text	lemmatized_text
0	As U.S. budget fight looms, Republicans flip their fiscal script WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansio...	1	as us budget fight looms republicans flip their fiscal script washington reuters the head of a conservative republican faction in the us congress who voted this month for a huge expansion of the ...	budget fight script head faction month expansion debt tax cut conservative budget restraint pivot way mark meadow nation line spending lawmaker battle holiday lawmaker budget fight issue immigrati...
1	U.S. military to accept transgender recruits on Monday: Pentagon WASHINGTON (Reuters) - Transgender people will be allowed for the first time to enlist in the U.S. military starting on Monday as o...	1	us military to accept transgender recruits on monday pentagon washington reuters transgender people will be allowed for the first time to enlist in the us military starting on monday as ordered b...	military transgender recruit people time military court administration ruling transgender ban appeal court week administration request hold order court judge military transgender recruit official ...
2	Senior U.S. Republican senator: 'Let Mr. Mueller do his job' WASHINGTON (Reuters) - The special counsel investigation of links between Russia and President Trump's 2016 election campaign should co...	1	senior us republican senator let mr mueller do his job washington reuters the special counsel investigation of links between russia and president trump's election campaign should continue withou...	job counsel investigation link trump election campaign interference call trump administration ally lawmaker senator force judiciary committee counsel investigation interference investigation inves...
3	FBI Russia probe helped by Australian diplomat tip-off: NYT WASHINGTON (Reuters) - Trump campaign adviser George Papadopoulos told an Australian diplomat in May 2016 that Russia had political dirt...	1	fbi russia probe helped by australian diplomat tipoff nyt washington reuters trump campaign adviser george papadopoulos told an australian diplomat in may that russia had political dirt on democ...	probe diplomat trump campaign adviser diplomat dirt candidate conversation papadopoulo diplomat alexander downer factor decision counterintelligence investigation contact trump campaign time month...
4	Trump wants Postal Service to charge 'much more' for Amazon shipments SEATTLE/WASHINGTON (Reuters) - President Donald Trump called on the U.S. Postal Service on Friday to charge "much more" to shi...	1	trump wants postal service to charge much more for amazon shipments seattlewashington reuters president donald trump called on the us postal service on friday to charge "much more" to ship packag...	trump service amazon shipment service ship package amazon amzno fight giant past office billion dollar year amazon package amazon post office trump twitter president tweet attention finance servic...

[26]:
```
# Check the dimensions of the DataFrame
df_clean.shape
```

[26]:
```
(44898, 4)
```

[27]:
```
# Check the dimensions of the DataFrame
df_clean.shape
```

[27]:
```
(44898, 4)
```

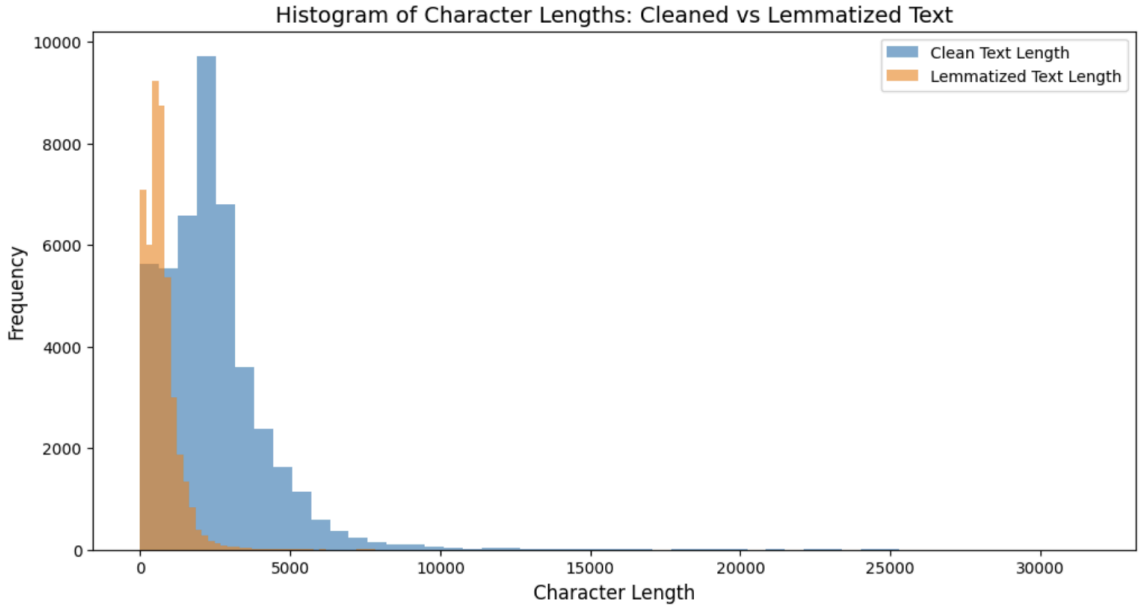
3. Train Validation Split

```python
# Import Train Test Split and split the DataFrame into 70% train and 30% validation data
from sklearn.model_selection import train_test_split

# Features and labels
X = df_clean['lemmatized_text']
y = df_clean['news_label']

# Train-validation split: 70% train, 30% validation
X_train, X_val, y_train, y_val = train_test_split(
    X, y, test_size=0.30, random_state=42, shuffle=True
)

# Check the shapes
print("Training set size:", X_train.shape[0])
print("Validation set size:", X_val.shape[0])
```

```
Training set size: 31428
Validation set size: 13470
```

4. Exploratory Data Analysis on Training Data

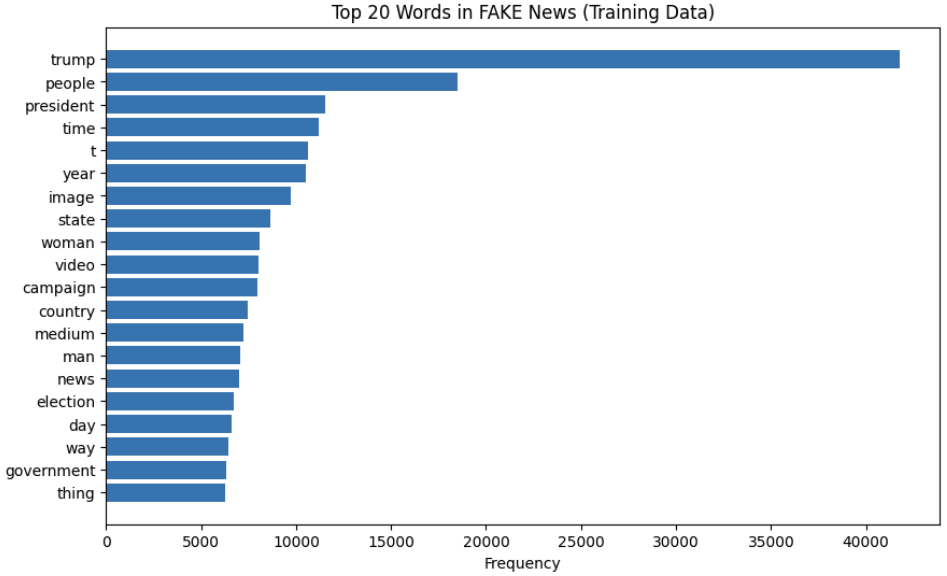
Visualised character lengths of cleaned news text and lemmatized news text with POS tags removed.




Top 40 Words in TRUE News (Training Data)
year
people
official
time
effort
decision
reuter
leader
attack
power
deal
policy
office
way
plan
day
state
issue
statement
trump
authority
administration
country
law
group
party
campaign
support
election
lawmaker
source
report
government
rule
member
case
president
week
month
company

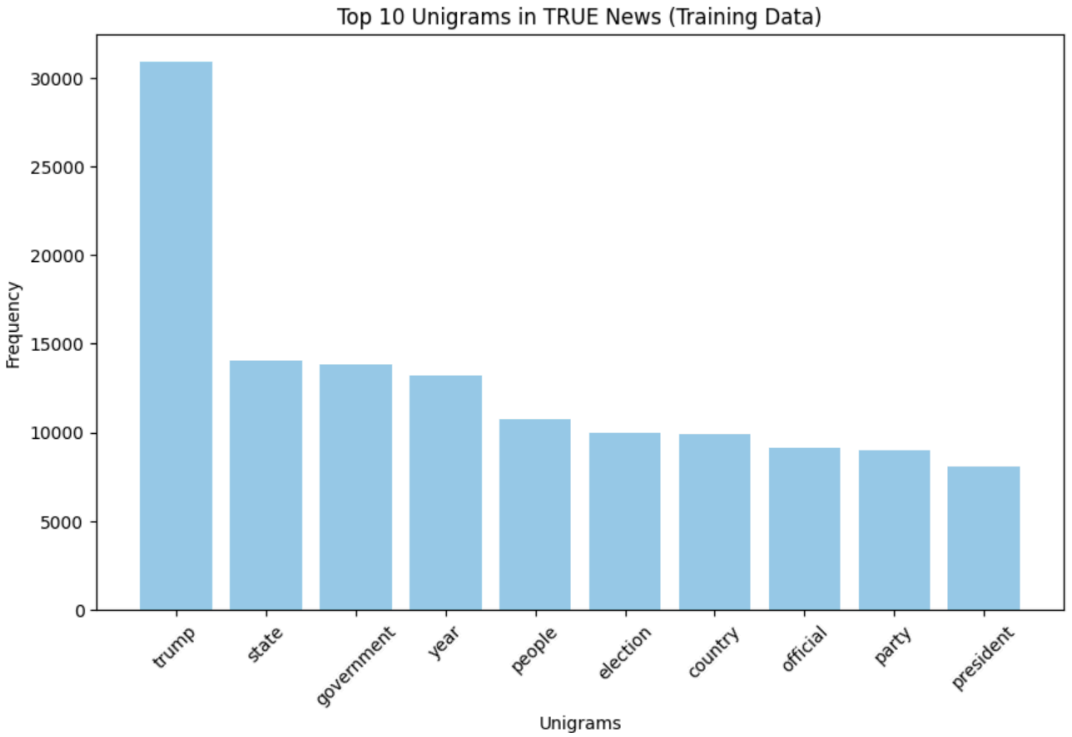
Top 40 Words in FAKE News (Training Data)

Top 20 Words in FAKE News (Training Data)



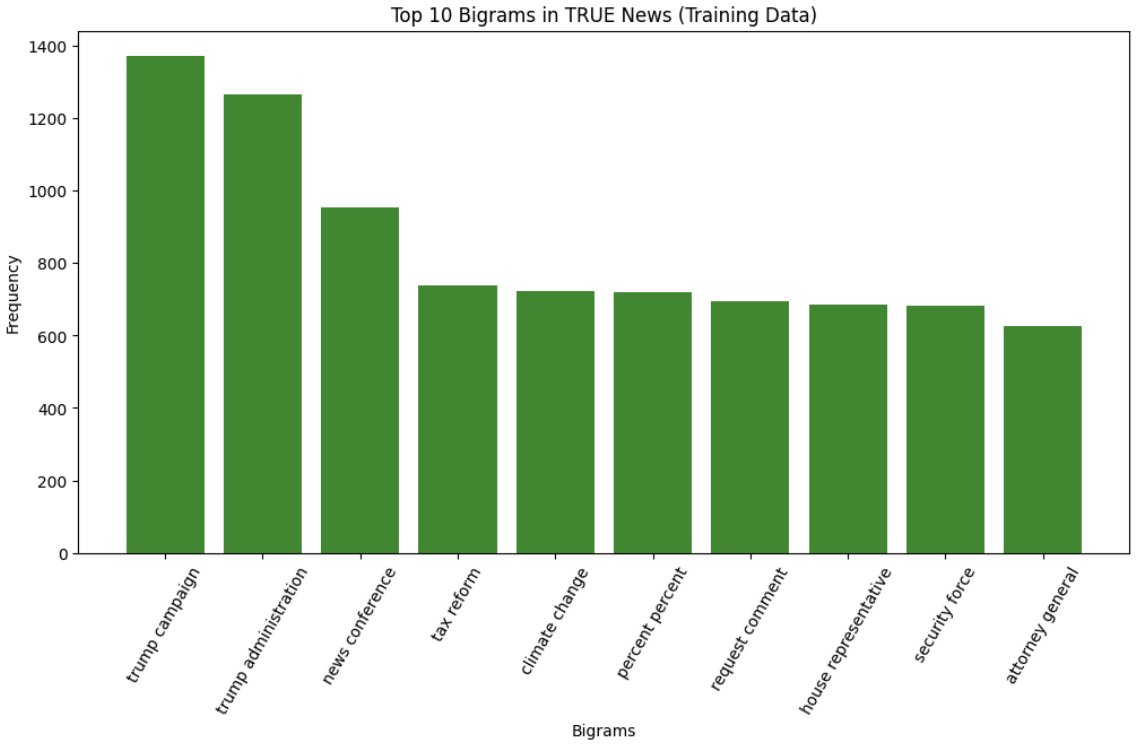
Top 10 Unigrams in TRUE News:

	ngram	frequency
0	trump	30893
1	state	14058
2	government	13846
3	year	13187
4	people	10761
5	election	9967
6	country	9902
7	official	9148
8	party	9016
9	president	8054



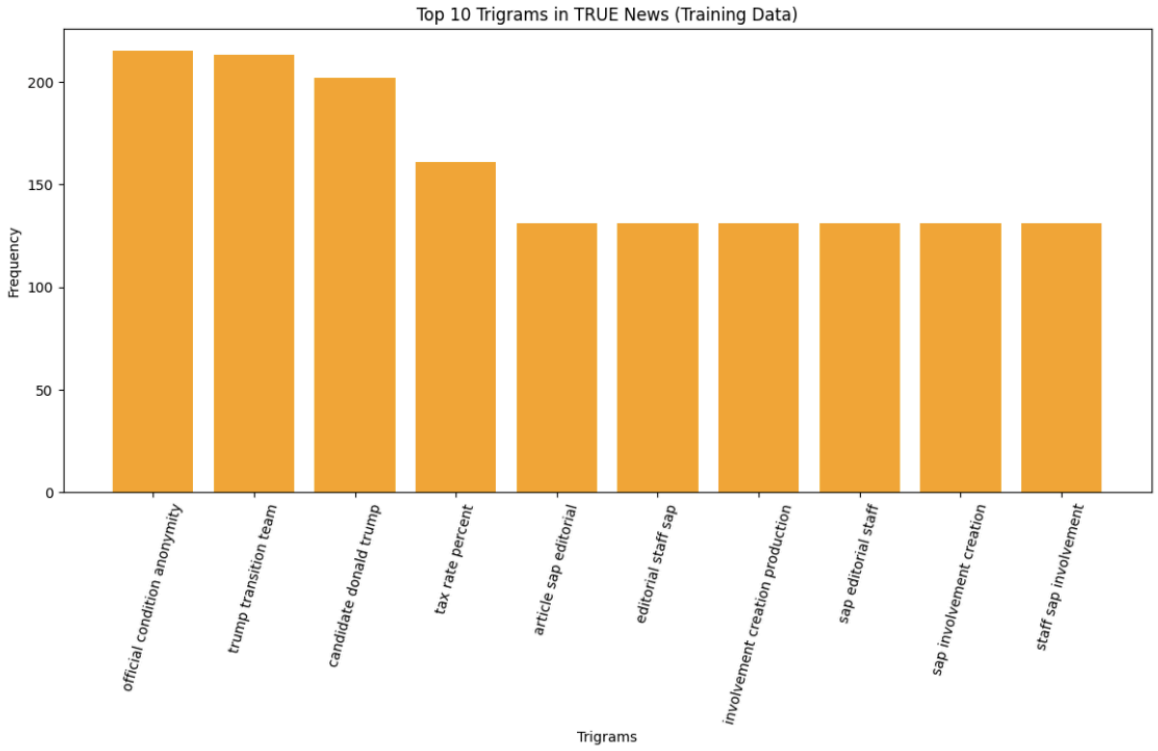
Top 10 Bigrams in TRUE News:

	ngram	frequency
0	trump campaign	1373
1	trump administration	1265
2	news conference	953
3	tax reform	740
4	climate change	722
5	percent percent	720
6	request comment	696
7	house representative	686
8	security force	683
9	attorney general	625



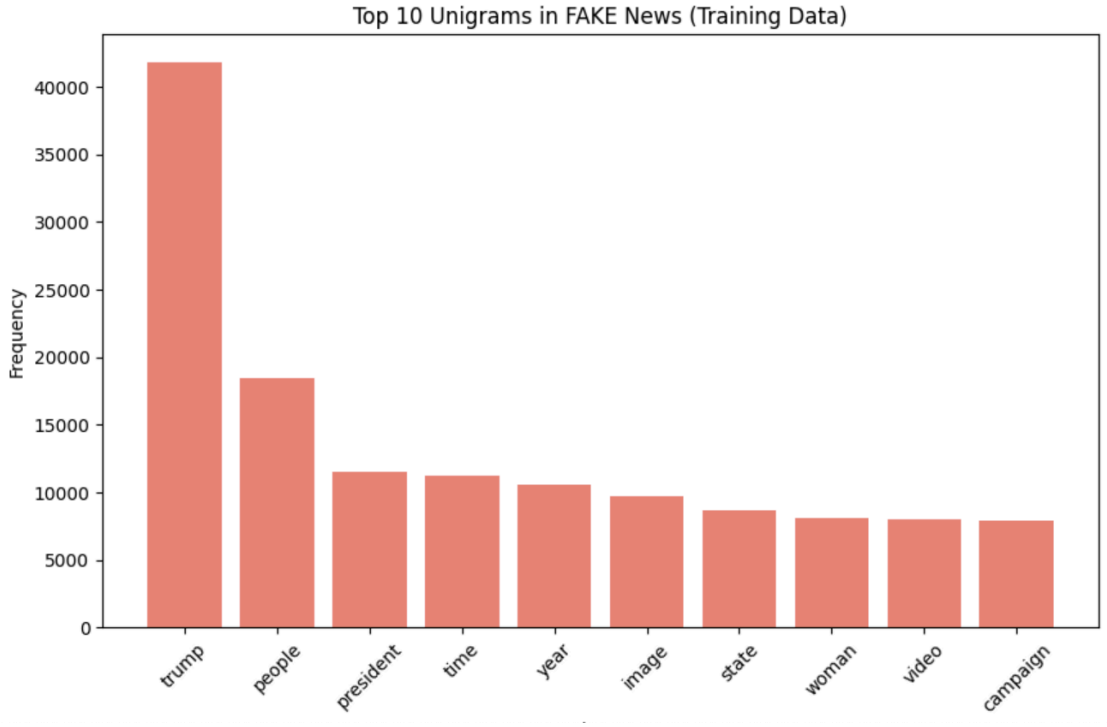
Top 10 Trigrams in TRUE News:

	ngram	frequency
0	official condition anonymity	215
1	trump transition team	213
2	candidate donald trump	202
3	tax rate percent	161
4	article sap editorial	131
5	editorial staff sap	131
6	involvement creation production	131
7	sap editorial staff	131
8	sap involvement creation	131
9	staff sap involvement	131



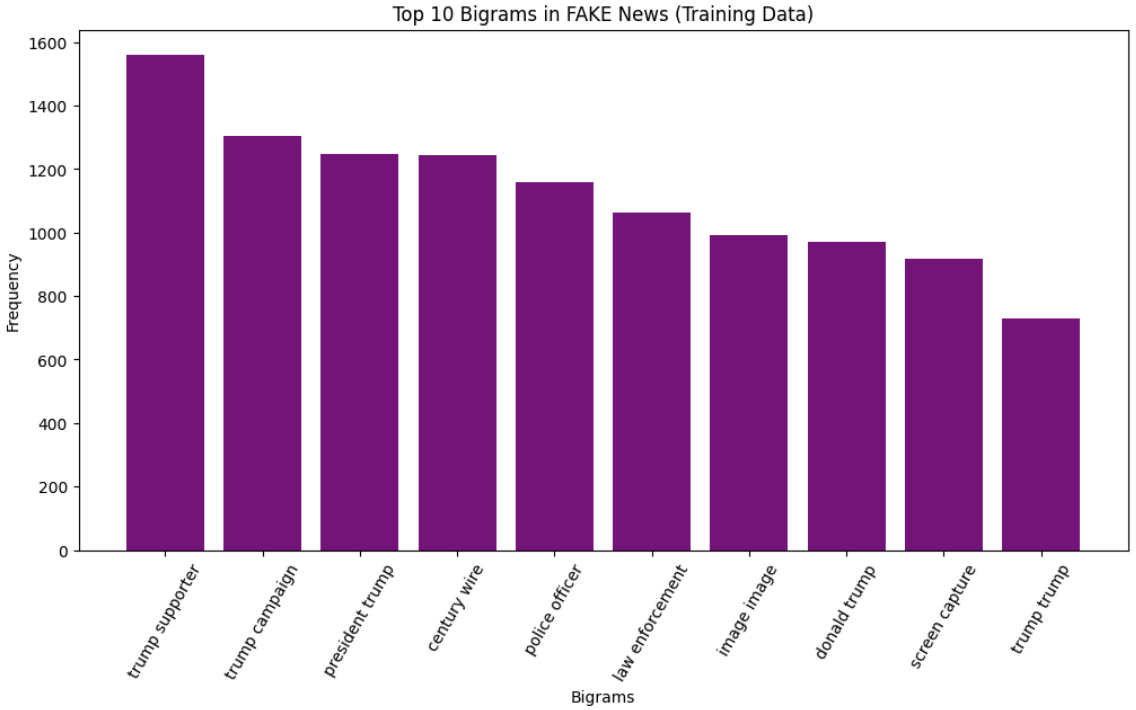
Top 10 Unigrams in FAKE News:

	ngram	frequency
0	trump	41801
1	people	18485
2	president	11515
3	time	11204
4	year	10529
5	image	9738
6	state	8653
7	woman	8070
8	video	8042
9	campaign	7939



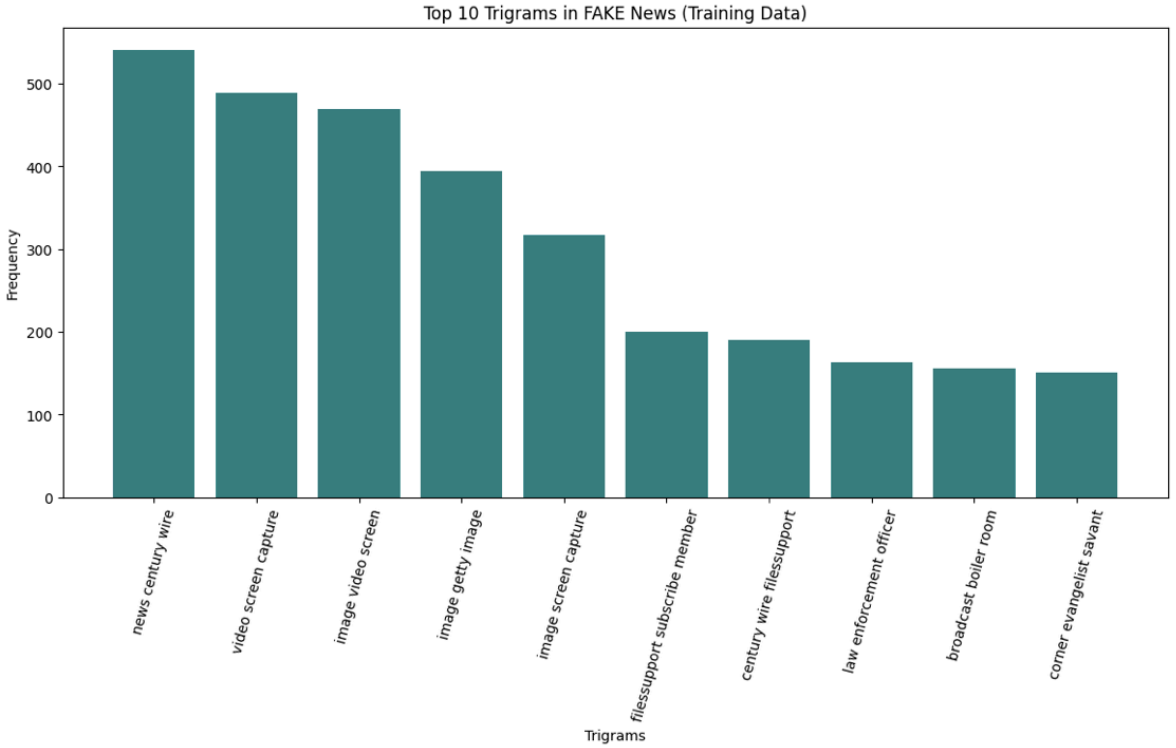
Top 10 Bigrams in FAKE News:

	ngram	frequency
0	trump supporter	1561
1	trump campaign	1305
2	president trump	1246
3	century wire	1245
4	police officer	1160
5	law enforcement	1064
6	image image	991
7	donald trump	970
8	screen capture	918
9	trump trump	729



Top 10 Trigrams in FAKE News:

	ngram	frequency
0	news century wire	541
1	video screen capture	489
2	image video screen	470
3	image getty image	394
4	image screen capture	317
5	filessupport subscribe member	200
6	century wire filessupport	190
7	law enforcement officer	163
8	broadcast boiler room	156
9	corner evangelist savant	151



6. Feature Extraction

Extract vectors for cleaned news data

```python
# 5) Compute vectors for train and val with progress bars
print("Building train vectors...")
X_train_vec = np.vstack([document_vector(txt, word2vec_model, vector_dim) for txt in tqdm(X_train.tolist())])

print("Building validation vectors...")
X_val_vec = np.vstack([document_vector(txt, word2vec_model, vector_dim) for txt in tqdm(X_val.tolist())])

# 6) Extract / ensure target variables are numpy arrays
y_train_arr = y_train.values if hasattr(y_train, "values") else np.array(y_train)
y_val_arr   = y_val.values   if hasattr(y_val, "values")   else np.array(y_val)

# 7) Quick sanity prints
print("X_train_vec shape:", X_train_vec.shape)
print("X_val_vec shape:  ", X_val_vec.shape)
print("y_train shape:", y_train_arr.shape)
print("y_val shape:  ", y_val_arr.shape)

# 8) Save vectors to disk (recommended to avoid re-computing)
np.save("X_train_vec.npy", X_train_vec)
np.save("X_val_vec.npy", X_val_vec)
np.save("y_train.npy", y_train_arr)
np.save("y_val.npy", y_val_arr)
print("Saved X_train_vec.npy, X_val_vec.npy, y_train.npy, y_val.npy")
```

```
Word2Vec vector dimension: 300
Building train vectors...
100%|██████████| 31428/31428 [00:11<00:00, 2742.56it/s]
Building validation vectors...
100%|██████████| 13470/13470 [00:03<00:00, 3961.74it/s]
X_train_vec shape: (31428, 300)
X_val_vec shape:   (13470, 300)
y_train shape: (31428,)
y_val shape:   (13470,)
Saved X_train_vec.npy, X_val_vec.npy, y_train.npy, y_val.npy
```

7. Model Training and Evaluation

7.1 Logistic Regression Model

-Created and trained logistic regression model on training data

-Calculated and printed accuracy, precision, recall and f1-score on validation data

```
Logistic Regression Evaluation Metrics:
---------------------------------------
Accuracy : 0.9014847809948032
Precision: 0.8962278675904541
Recall   : 0.8991350015446401
F1-score : 0.8976790808851878

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.90      0.91      6996
           1       0.90      0.90      0.90      6474

    accuracy                           0.90     13470
   macro avg       0.90      0.90      0.90     13470
weighted avg       0.90      0.90      0.90     13470
```

]:
```python
# Classification Report

# Classification Report for Logistic Regression

print("Classification Report (Logistic Regression):")
print("-------------------------------------------")
print(classification_report(y_val_arr, y_pred_logreg))
```

```
Classification Report (Logistic Regression):
--------------------------------------------
              precision    recall  f1-score   support

           0       0.91      0.90      0.91      6996
           1       0.90      0.90      0.90      6474

    accuracy                           0.90     13470
   macro avg       0.90      0.90      0.90     13470
weighted avg       0.90      0.90      0.90     13470
```

7.2 Decision Tree Model

-Created and trained a decision tree model on training data

-Calculated and printed accuracy, precision, recall and f1-score on validation data

```python
# Calculate and print accuracy, precision, recall, f1-score on predicted labels

accuracy_dt = accuracy_score(y_val_arr, y_pred_dt)
precision_dt = precision_score(y_val_arr, y_pred_dt)
recall_dt = recall_score(y_val_arr, y_pred_dt)
f1_dt = f1_score(y_val_arr, y_pred_dt)

print("Decision Tree Evaluation Metrics:")
print("---------------------------------")
print("Accuracy :", accuracy_dt)
print("Precision:", precision_dt)
print("Recall   :", recall_dt)
print("F1-score :", f1_dt)
```

```
Decision Tree Evaluation Metrics:
---------------------------------
Accuracy : 0.8233853006681514
Precision: 0.8352710004912396
Recall   : 0.7879209144269386
F1-score : 0.8109053334393133
```

```python
# Classification Report

print("\nClassification Report:")
print(classification_report(y_val_arr, y_pred_dt))
```

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

           0       0.81      0.86      0.83      6996
           1       0.84      0.79      0.81      6474

    accuracy                           0.82     13470
   macro avg       0.82      0.82      0.82     13470
weighted avg       0.82      0.82      0.82     13470
```

7.3 Random Forest Model

-Created and trained a random forest model on training data

-Calculated and printed accuracy, precision, recall and f1-score on validation data

```
## Calculate and print accuracy, precision, recall, f1-score on predicted labels

accuracy_rf = accuracy_score(y_val_arr, y_pred_rf)
precision_rf = precision_score(y_val_arr, y_pred_rf)
recall_rf = recall_score(y_val_arr, y_pred_rf)
f1_rf = f1_score(y_val_arr, y_pred_rf)

print("Random Forest Evaluation Metrics:")
print("---------------------------------")
print("Accuracy :", accuracy_rf)
print("Precision:", precision_rf)
print("Recall   :", recall_rf)
print("F1-score :", f1_rf)
```

```
Random Forest Evaluation Metrics:
---------------------------------
Accuracy : 0.9074981440237565
Precision: 0.9113943972300913
Recall   : 0.8945010812480692
F1-score : 0.9028687246647957
```

```
# Classification Report

print("\nClassification Report:")
print(classification_report(y_val_arr, y_pred_rf))
```

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

           0       0.90      0.92      0.91      6996
           1       0.91      0.89      0.90      6474

    accuracy                           0.91     13470
   macro avg       0.91      0.91      0.91     13470
weighted avg       0.91      0.91      0.91     13470
```

8. Conclusion

In this project, we cleaned and processed news text, removed unnecessary words, and kept only important nouns to better understand the meaning.

We observed that:

- True news tended to use more factual and specific nouns (such as government, report, official).
- Fake news often repeated sensational or dramatic nouns (like claim, video, story).

Using Word2Vec helped capture the semantic meaning of words instead of just counting them. Similar words ended up with similar vectors, which allowed the model to better understand context.

We trained three models — Logistic Regression, Decision Tree, and Random Forest.

Among them, Random Forest performed the best, especially in terms of F1-score, which is important because:

- We don’t want to incorrectly label fake news as true.
- We also shouldn’t label true news as fake.

Overall, the semantic approach (cleaning + lemmatization + Word2Vec) significantly improved the model’s ability to recognize patterns and detect fake news. The project highlighted clear differences between true and fake news, and Random Forest’s strong F1-score made it the most reliable model for this task.