2025

Case Study Report Fake News Detection using Semantic Classification

By: Kafeel Ahmad and Isha Khatkar

UpGrad

6/22/2025

Contents

Fake News Detection Using Semantic Classification

1. Problem Statement

Fake news poses a significant threat to public discourse by spreading misinformation across digital platforms. The purpose of this project is to develop a semantic classification model using Word2Vec embeddings to accurately differentiate between fake and true news articles. This approach focuses not just on keywords but on capturing the contextual meaning of the news content.

2. 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.

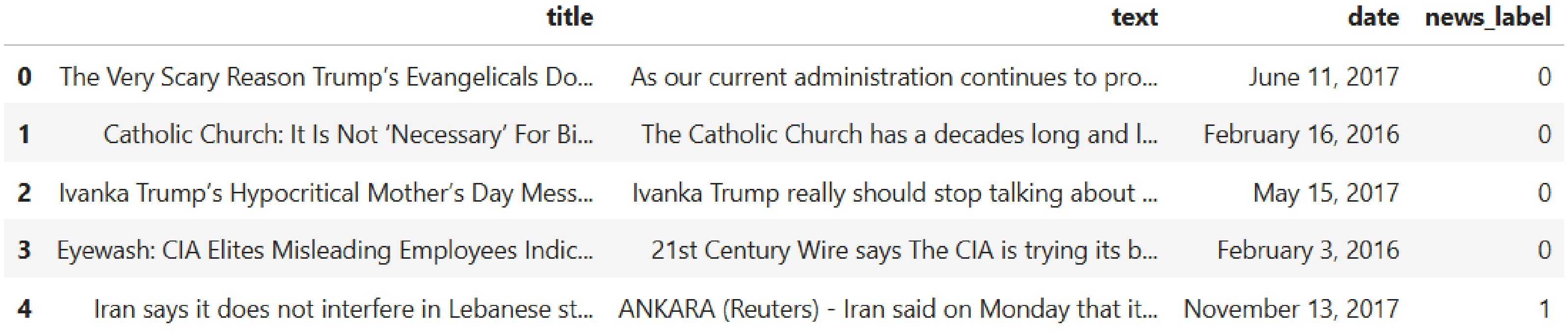
3. Methodology and Techniques

The following steps were used to build the fake news classification system:

- Data Loading: Two datasets containing true and fake news articles were combined.
- Preprocessing: Cleaning, lemmatization, and POS tagging were performed to prepare the data.
- Feature Extraction: Word2Vec embeddings (Google News 300 dimensions) were used to represent news text semantically.
- Train-Validation Split: The dataset was split into 70% training and 30% validation sets.
- Exploratory Data Analysis: Character lengths, word clouds, and n-gram analysis were used to extract insights.
- Model Training: Logistic Regression, Decision Tree, and Random Forest classifiers were trained and evaluated.
- Evaluation: Models were assessed using Accuracy, Precision, Recall, and F1-Score.

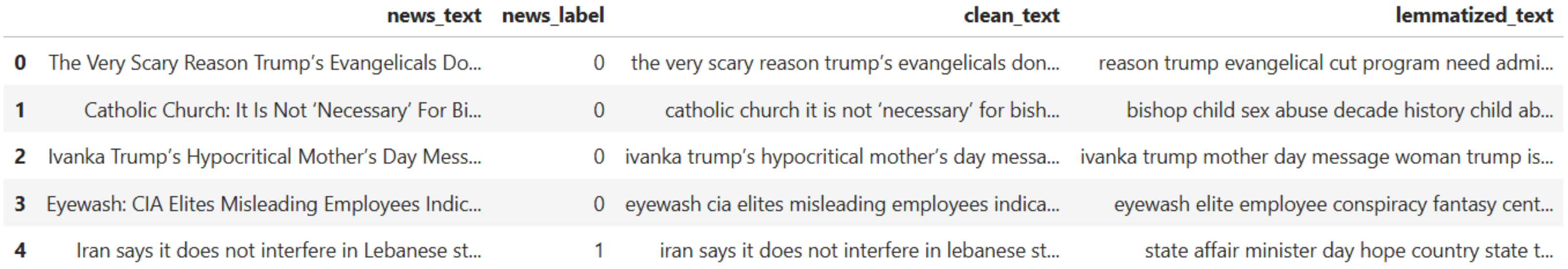
4. Exploratory Data Analysis & Visualizations

4.1. Sample Data:



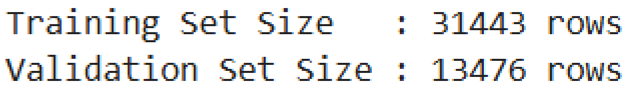
	title	text	date	news_label
0	The Very Scary Reason Trump's Evangelicals Do...	As our current administration continues to pro...	June 11, 2017	0
1	Catholic Church: It Is Not 'Necessary' For Bi...	The Catholic Church has a decades long and l...	February 16, 2016	0
2	Ivanka Trump's Hypocritical Mother's Day Mess...	Ivanka Trump really should stop talking about ...	May 15, 2017	0
3	Eyewash: CIA Elites Misleading Employees Indic...	21st Century Wire says The CIA is trying its b...	February 3, 2016	0
4	Iran says it does not interfere in Lebanese st...	ANKARA (Reuters) - Iran said on Monday that it...	November 13, 2017	1

4.2. Clean and Lemmatized Text with shape of (44919, 4):



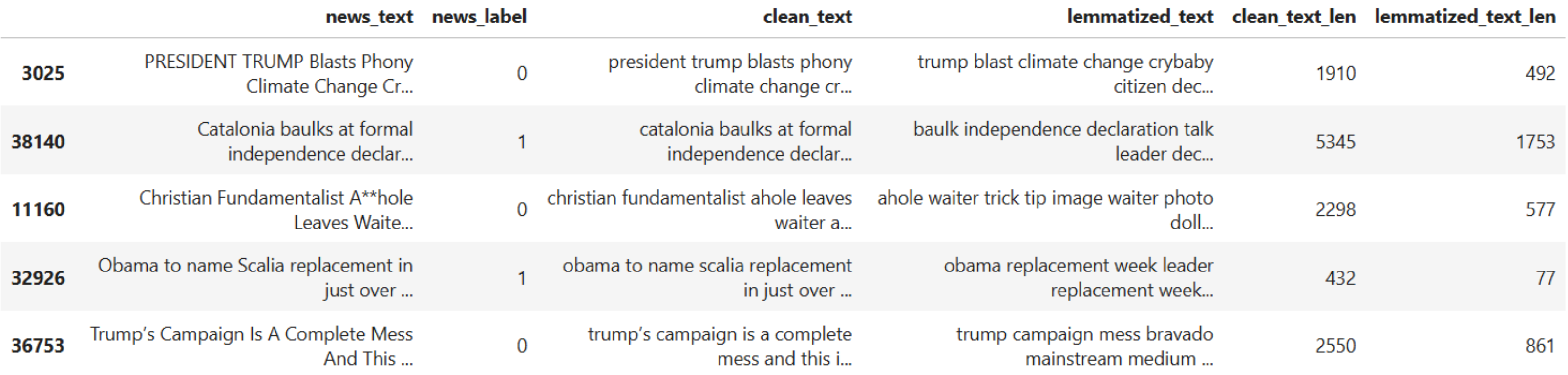
	news_text	news_label	clean_text	lemmatized_text
0	The Very Scary Reason Trump's Evangelicals Do...	0	the very scary reason trump's evangelicals don...	reason trump evangelical cut program need admi...
1	Catholic Church: It Is Not 'Necessary' For Bi...	0	catholic church it is not 'necessary' for bish...	bishop child sex abuse decade history child ab...
2	Ivanka Trump's Hypocritical Mother's Day Mess...	0	ivanka trump's hypocritical mother's day messa...	ivanka trump mother day message woman trump is...
3	Eyewash: CIA Elites Misleading Employees Indic...	0	eyewash cia elites misleading employees indica...	eyewash elite employee conspiracy fantasy cent...
4	Iran says it does not interfere in Lebanese st...	1	iran says it does not interfere in lebanese st...	state affair minister day hope country state t...

4.3. Train Validation Split:



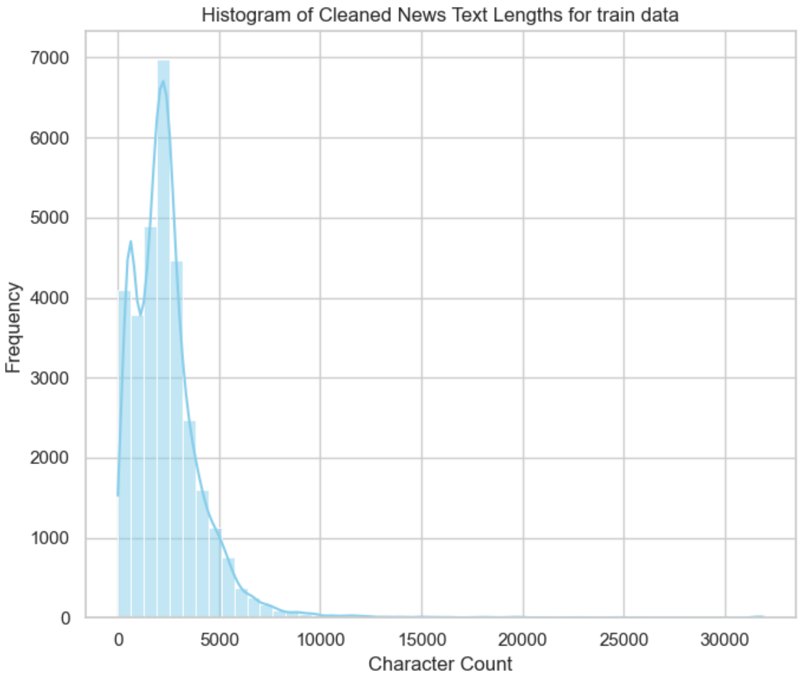
```
Training Set Size   : 31443 rows
Validation Set Size : 13476 rows
```

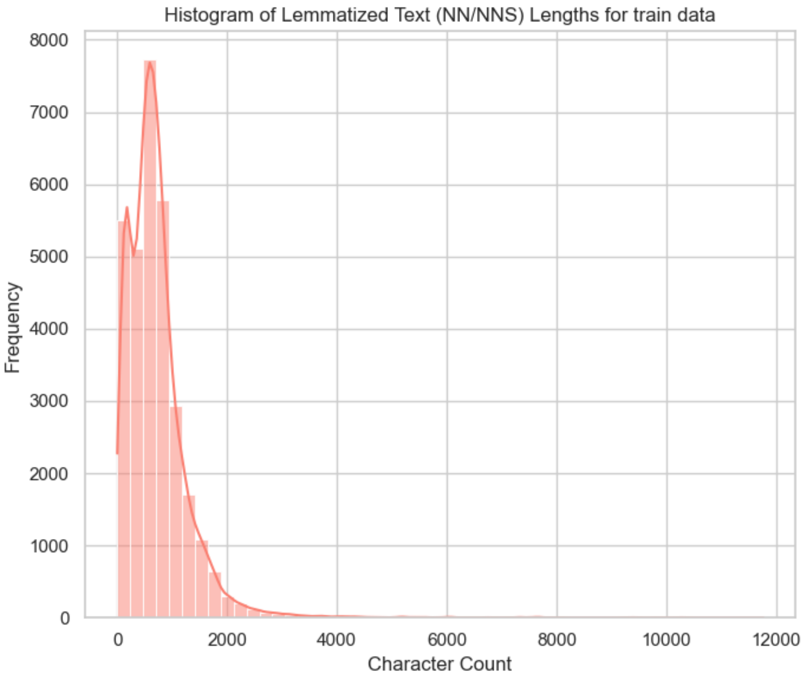
4.4. For EDA and Visualization Clean and Lemmatized text length calculated:



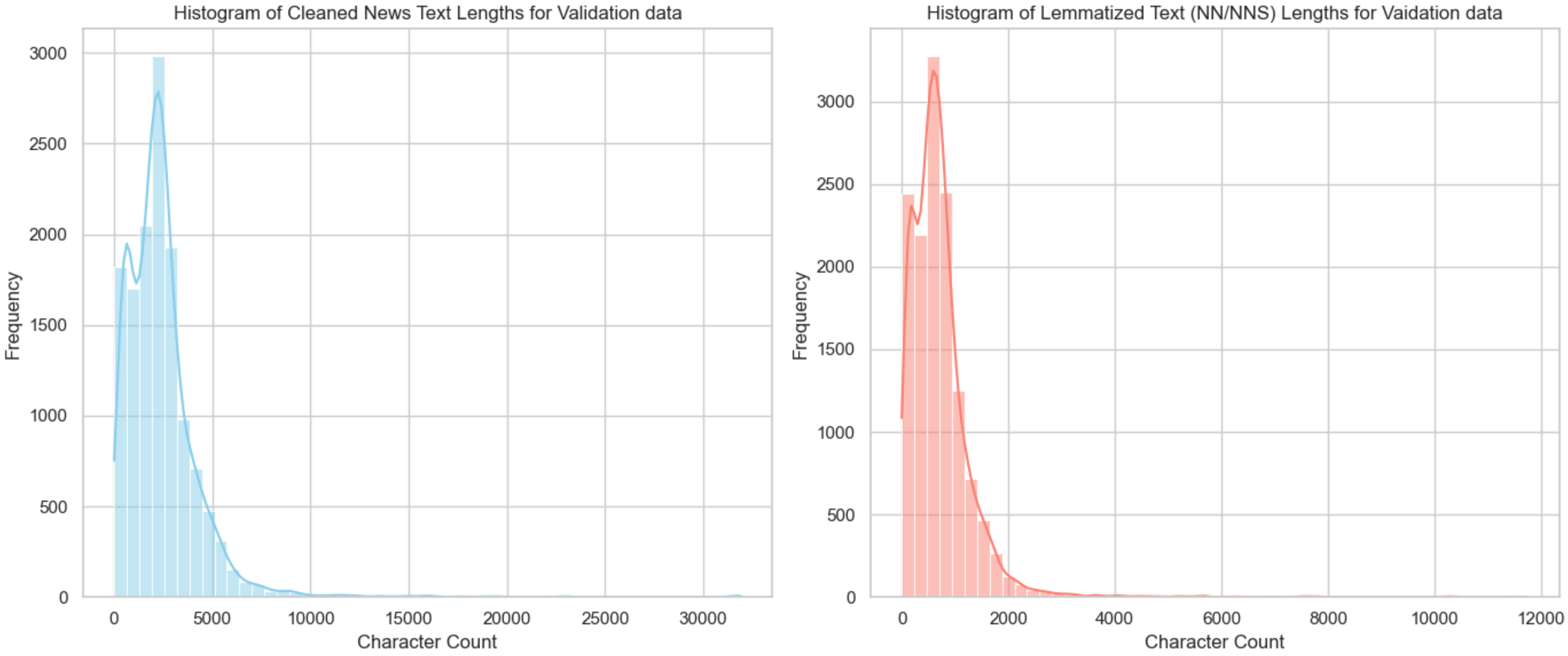
	news_text	news_label	clean_text	lemmatized_text	clean_text_len	lemmatized_text_len
3025	PRESIDENT TRUMP Blasts Phony Climate Change Cr...	0	president trump blasts phony climate change cr...	trump blast climate change crybaby citizen dec...	1910	492
38140	Catalonia baulks at formal independence declar...	1	catalonia baulks at formal independence declar...	baulk independence declaration talk leader dec...	5345	1753
11160	Christian Fundamentalist A**hole Leaves Waite...	0	christian fundamentalist ahole leaves waiter a...	ahole waiter trick tip image waiter photo doll...	2298	577
32926	Obama to name Scalia replacement in just over ...	1	obama to name scalia replacement in just over ...	obama replacement week leader replacement week...	432	77
36753	Trump’s Campaign Is A Complete Mess And This ...	0	trump’s campaign is a complete mess and this i...	trump campaign mess bravado mainstream medium ...	2550	861

4.5. Histogram of Cleaned and Lemmatized Text for Train Data

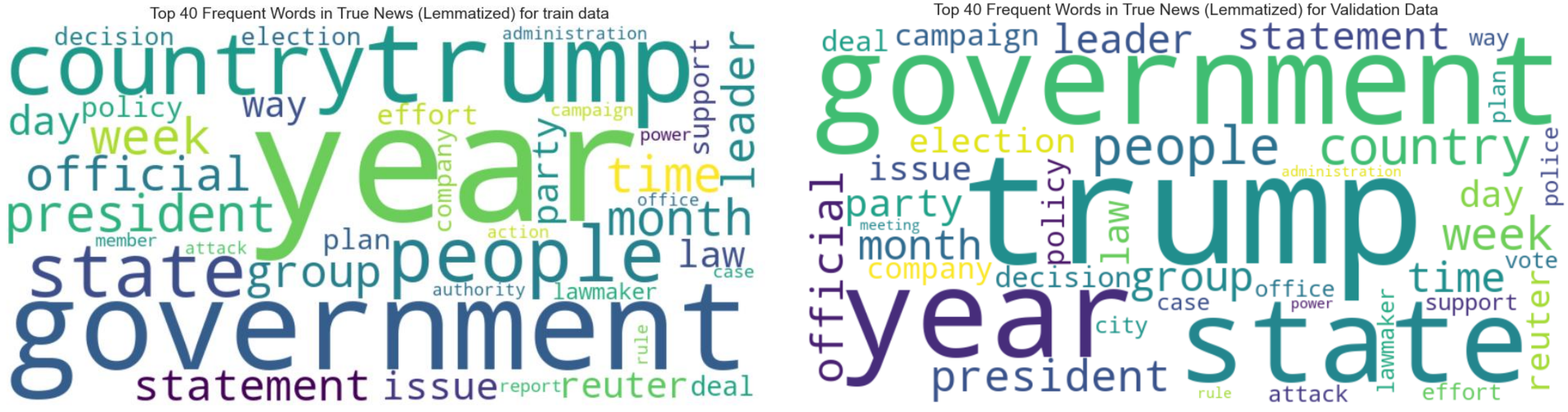




4.6. Histogram of Cleaned and Lemmatized Text for Validation Data

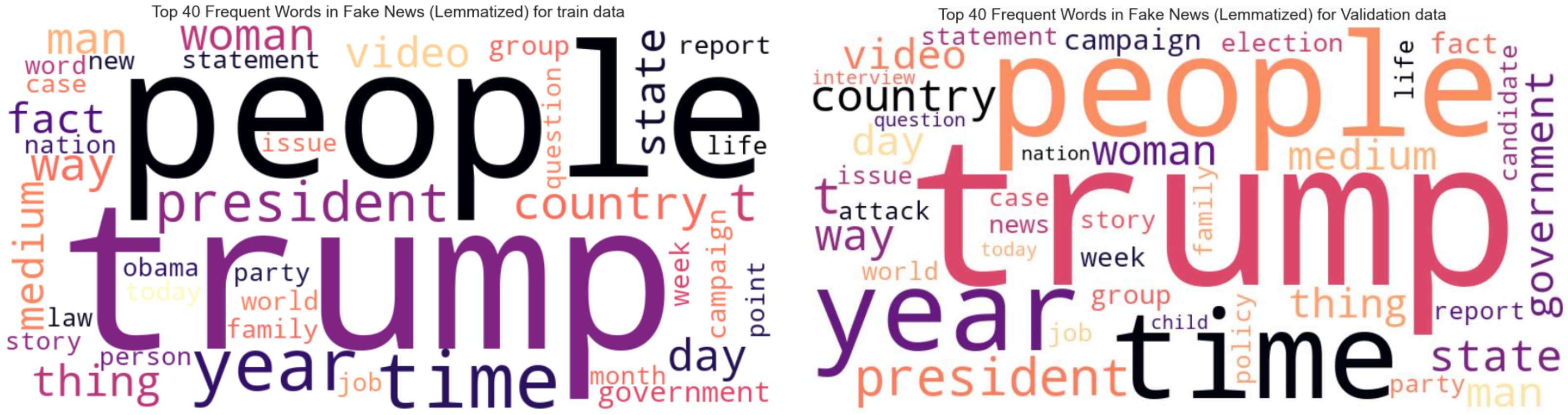


4.7. Top 40 Frequent Words in TRUE News (Lemmatized) for Train and Validation data


Top 40 Frequent Words in True News (Lemmatized) for train data

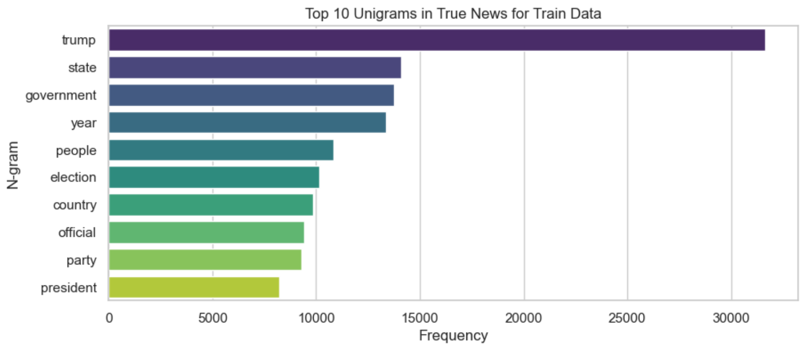
Top 40 Frequent Words in True News (Lemmatized) for Validation Data

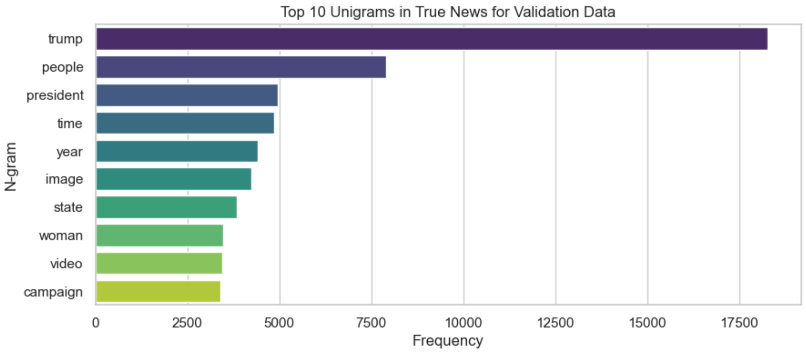
4.8. Top 40 Frequent Words in FAKE News (Lemmatized) for Train and Validation data


Top 40 Frequent Words in Fake News (Lemmatized) for train data

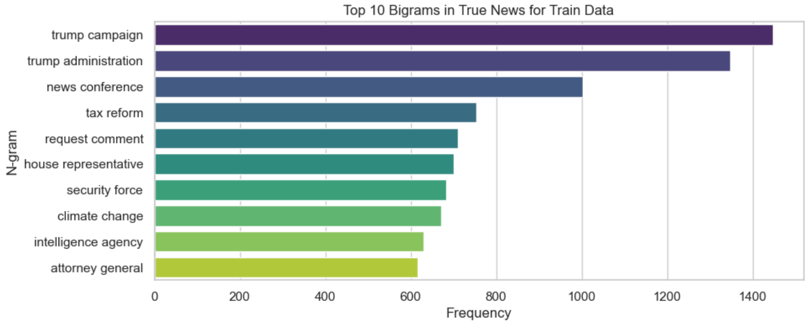
Top 40 Frequent Words in Fake News (Lemmatized) for Validation data

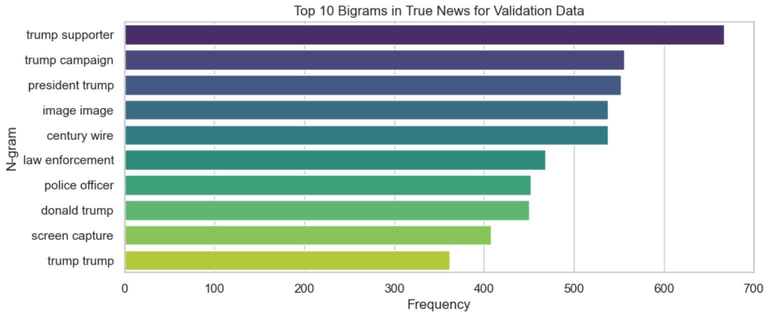
4.9. Top 10 Unigrams in TRUE News for Train and Validation data


Top 10 Unigrams in True News for Train Data

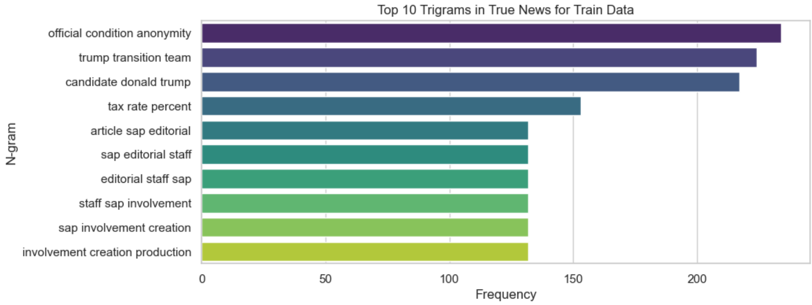

Top 10 Unigrams in True News for Validation Data

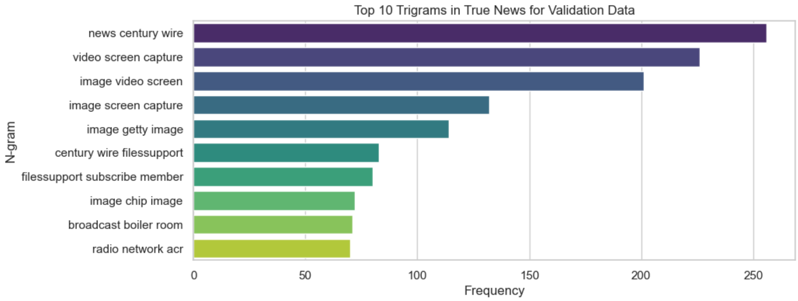
4.10. Top 10 Bigrams in TRUE News for Train and Validation data


Top 10 Bigrams in True News for Train Data

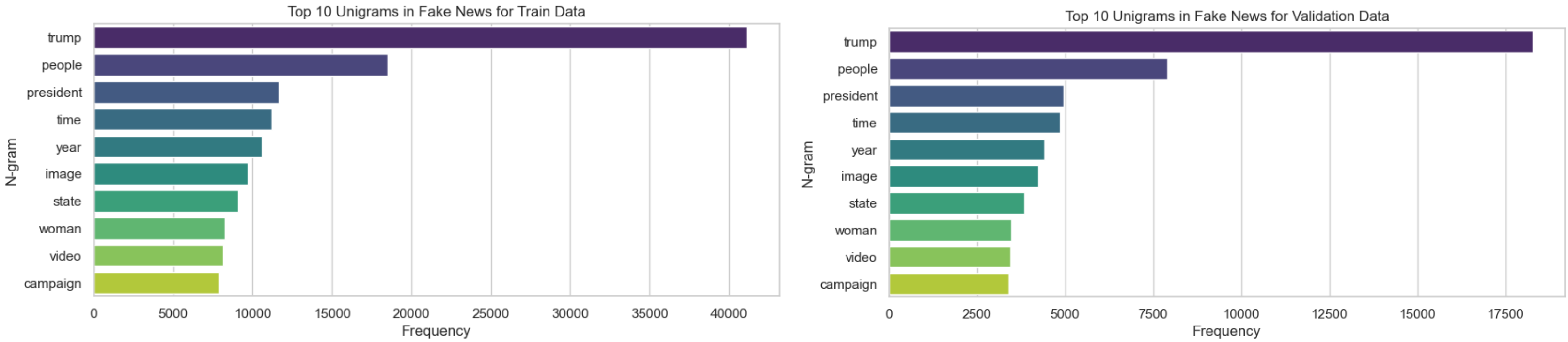

Top 10 Bigrams in True News for Validation Data

4.11. Top 10 Trigrams in TRUE News for Train and Validation data

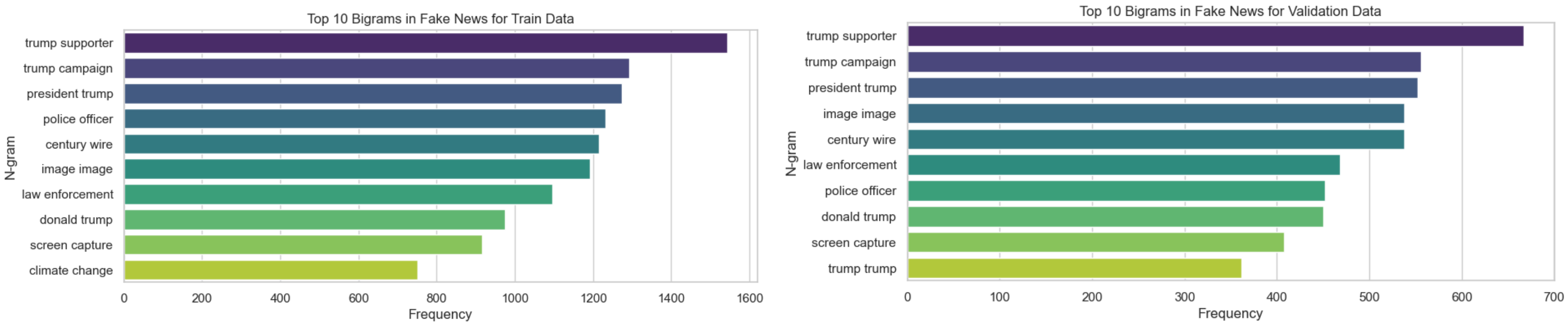

Top 10 Trigrams in True News for Train Data


Top 10 Trigrams in True News for Validation Data

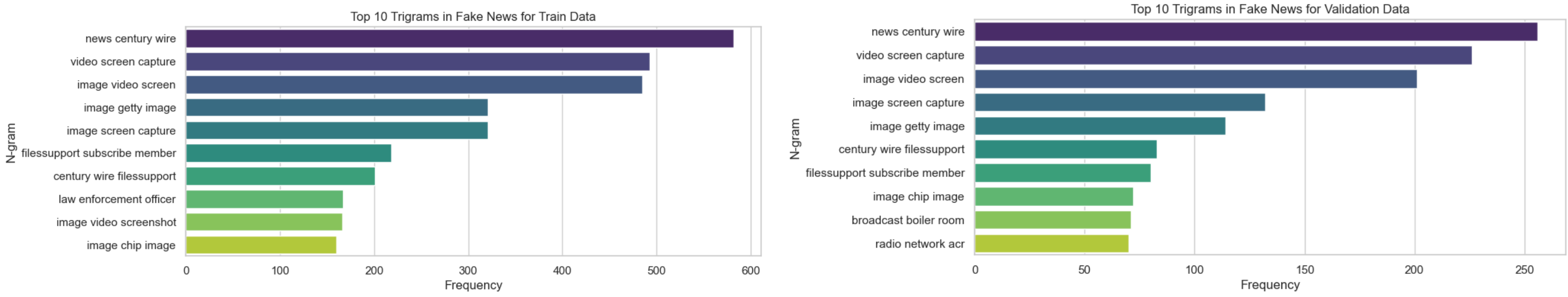
4.12. Top 10 Unigrams in FAKE News for Train and Validation data



4.13. Top 10 Bigrams in FAKE News for Train and Validation data



4.14. Top 10 Trigrams in FAKE News for Train and Validation data



5. Building Model

5.1. Logistic Regression Model:

Evaluation Metrics for Logistic Regression Model:

Accuracy	**0.9044**
Precision	**0.8982**
Recall	**0.9018**
F1 Score	**0.9000**

Classification Report for Logistic Regression Model:

	Precision	**Recall**	**F1 Score**	**Support**
0	0.91	0.91	0.91	7051
1	0.90	0.90	0.90	6425
Accuracy			0.90	13476
Macro Avg.	0.90	0.90	0.90	13476
Weighted Avg.	0.90	0.90	0.90	13476

5.2. Decision Tree Model:

Evaluation Metrics for Decision Tree Model:

Accuracy	**0.8263**
Precision	**0.8302**
Recall	**0.7991**
F1 Score	**0.8143**

Classification Report for Decision Tree Model:

	Precision	Recall	F1 Score	Support
0	0.82	0.85	0.84	7051
1	0.83	0.80	0.81	6425
Accuracy			0.83	13476
Macro Avg.	0.83	0.83	0.83	13476
Weighted Avg.	0.83	0.83	0.83	13476

5.3. Random Forest Model:

Evaluation Metrics for Random Forest Model:

Accuracy	**0.9085**
Precision	**0.9146**
Recall	**0.8914**
F1 Score	**0.9028**

Classification Report for Random Forest Model:

	Precision	Recall	F1 Score	Support
0	0.90	0.92	0.91	7051
1	0.91	0.89	0.90	6425
Accuracy			0.91	13476
Macro Avg.	0.91	0.91	0.91	13476
Weighted Avg.	0.91	0.91	0.91	13476

6. Model Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	90.44%	89.82%	**90.18%**	90.00%
Decision Tree	82.63%	83.02%	79.91%	81.43%
Random Forest	**90.85%**	**91.46%**	89.14%	**90.28%**

- **Best Model Chosen: Random Forest**
- **Evaluation Metric Prioritized: F1 Score**
  - Since this is a **binary classification** problem with potential cost on both false positives (labeling true news as fake) and false negatives (allowing fake news to pass as true), **F1 score provides** a balanced measure of precision and recall.

7. Insights and Analysis

The Word2Vec-based semantic embedding helped to extract meaningful relationships between words beyond simple keyword matching. EDA indicated that true news articles tend to use formal, topic-specific vocabulary, while fake news often uses emotionally charged or exaggerated terms. N-gram analysis and word clouds further reinforced this pattern.

Among all models, Random Forest performed the best with over 90% in all evaluation metrics. Logistic Regression was a close second. The Decision Tree model underperformed slightly due to overfitting tendencies without regularization.

8. Conclusion and Actionable Outcomes

This project demonstrates the effectiveness of semantic classification using Word2Vec embeddings in detecting fake news. The Random Forest model, trained on semantically enriched features, provided the most reliable performance.

Actionable outcomes from this work include:
- Integrating the trained model into a real-time news validation system

- Extending the model using contextual embeddings like BERT for even better performance
- Applying similar methodologies to other misinformation-sensitive domains such as medical or financial news

By focusing on the meaning and structure of the text rather than just surface-level keywords, this approach offers a scalable and effective solution to mitigating fake news propagation online.