FAKE NEWS DETECTION

Prepared by Amit Roy and Ankit Shrivastava

BUSINESS OBJECTIVE

The Challenge

- Problem: The spread of fake news has become a critical threat to informed decision-making and public trust
- Volume: Massive daily publication of news articles makes manual verification impossible
- Impact: Misinformation undermines democratic processes, public health decisions, and social cohesion
- Need: Automated systems capable of distinguishing credible from misleading information

Business Goals

- Develop an intelligent semantic classification system using Word2Vec to automatically identify fake news
- Achieve high accuracy while maintaining balance between precision and recall
- Create a scalable solution for real-time news verification based on semantic meaning rather than just syntax
- Provide interpretable results for content moderation decisions

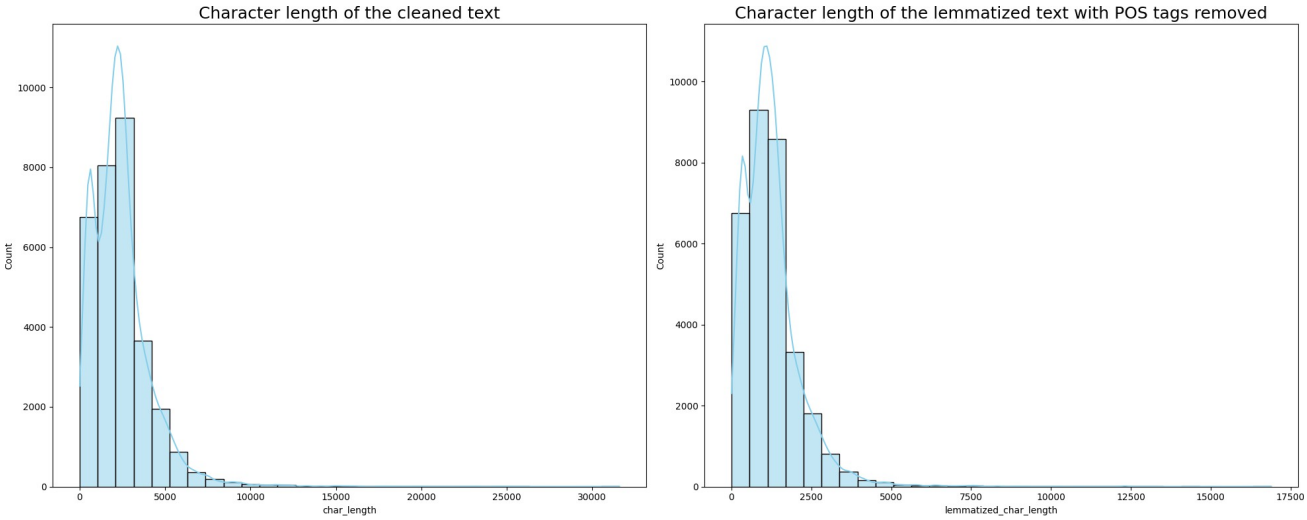
DATA OVERVIEW & EXPLORATION

Dataset Characteristics

- Training Set: 31,428 news articles
- Validation Set: 13,470 news articles
- Class Distribution: Well-balanced dataset (Fake News: 7,045 samples, True News: 6,425 samples in validation)
- Text Processing: Lemmatized content with POS tags removed for optimal semantic analysis

Character Length Analysis

- Cleaned Text: Distribution peaks around 8,000-10,000 characters with right- skewed pattern
- Lemmatized Text: Similar distribution pattern after POS tag removal, confirming effective preprocessing
- Consistency: Both distributions show comparable patterns, validating text processing pipeline



DATA PROCESSING APPROACH

Text Preprocessing Steps

Stage 1: Basic Text Cleaning

- Convert all text to lowercase for consistency
- Remove content within square brackets (citations, references)
- Strip all punctuation marks
- Eliminate words containing numbers (dates, statistics)

Stage 2: Linguistic Processing

- Tokenization: Split text into individual words
- POS Tagging: Identify grammatical roles of each word
- Noun Extraction: Filter and retain only nouns (NN, NNS tags) for semantic focus
- Stop Word Removal: Remove common English stop words (the, and, is, etc.)
- Lemmatization: Convert words to their root forms (running → run, better → good)

Stage 3: Word2Vec Preparation

- Format Cleaning: Remove brackets, quotes, and commas from processed text

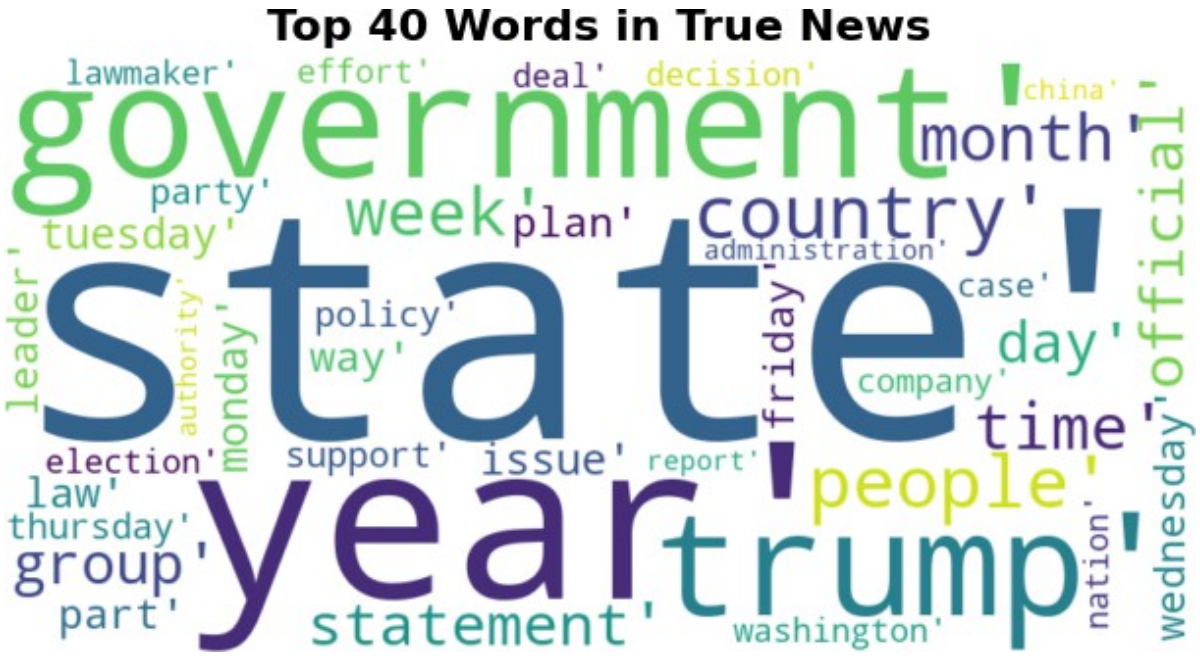
- Whitespace Normalization: Standardize spacing between words
- Word Splitting: Convert cleaned text into individual word tokens

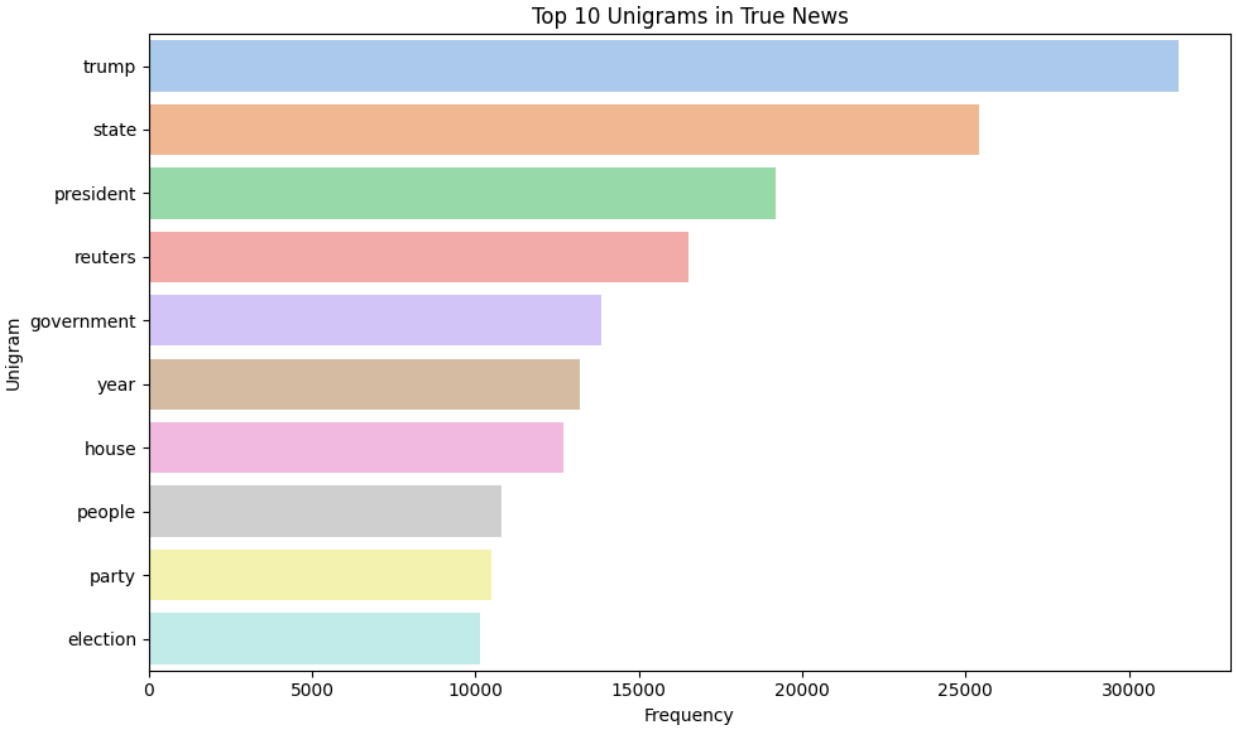
Stage 4: Semantic Vector Extraction

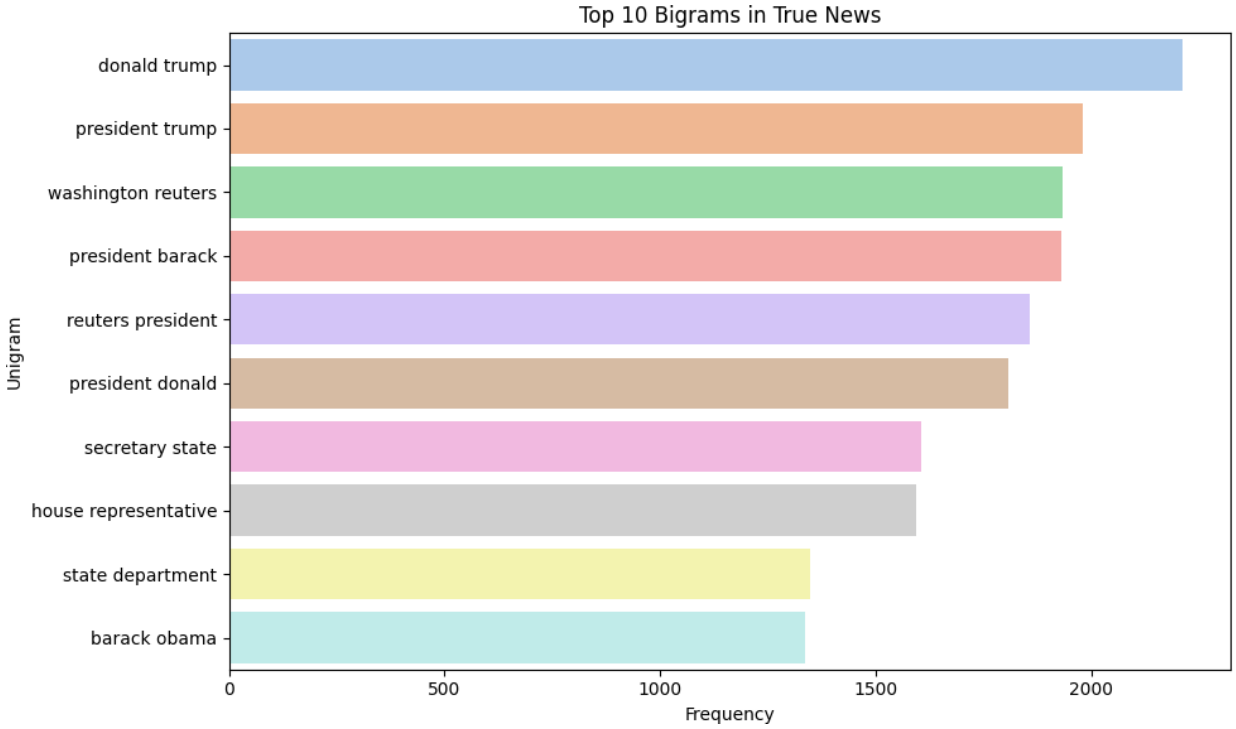
- Vocabulary Matching: Check each word against Word2Vec model vocabulary
- Vector Retrieval: Extract 300-dimensional vectors for matched words
- Document Vectorization: Average all word vectors to create single document vector
- Missing Handling: Generate zero vectors for texts with no vocabulary matches

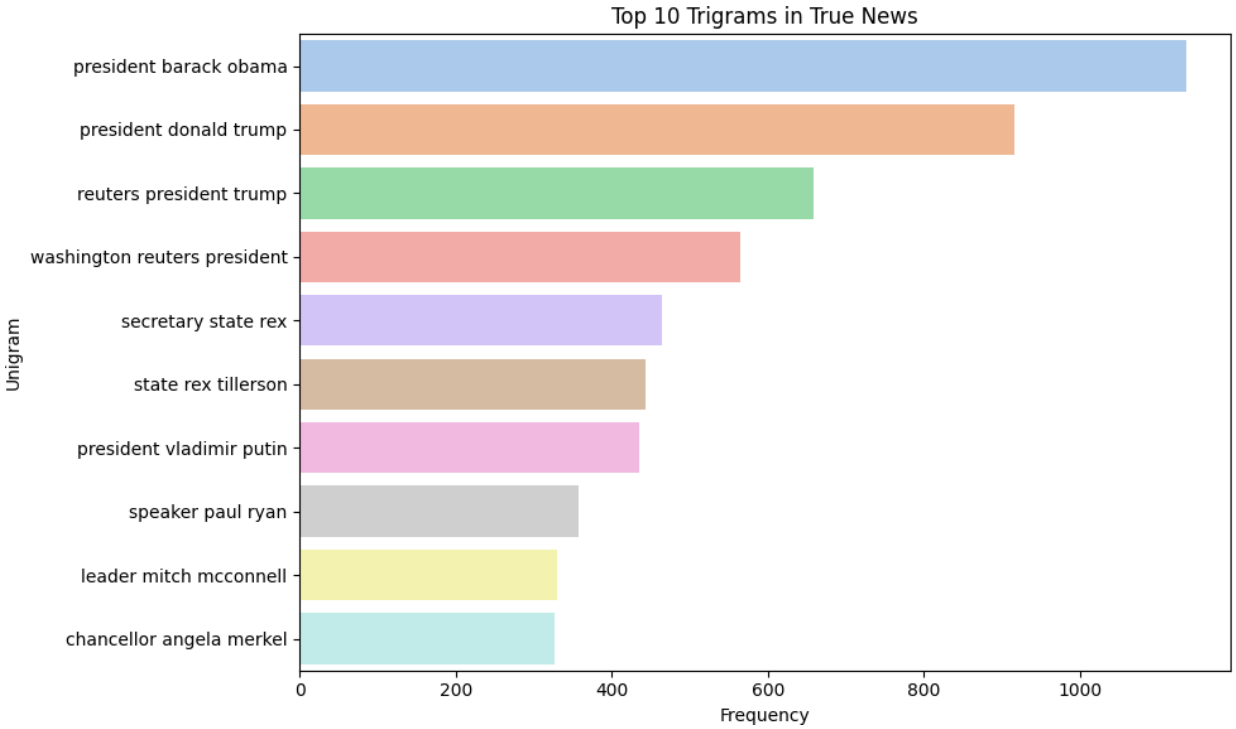
N-GRAM ANALYSIS FINDINGS

True News Linguistic Patterns

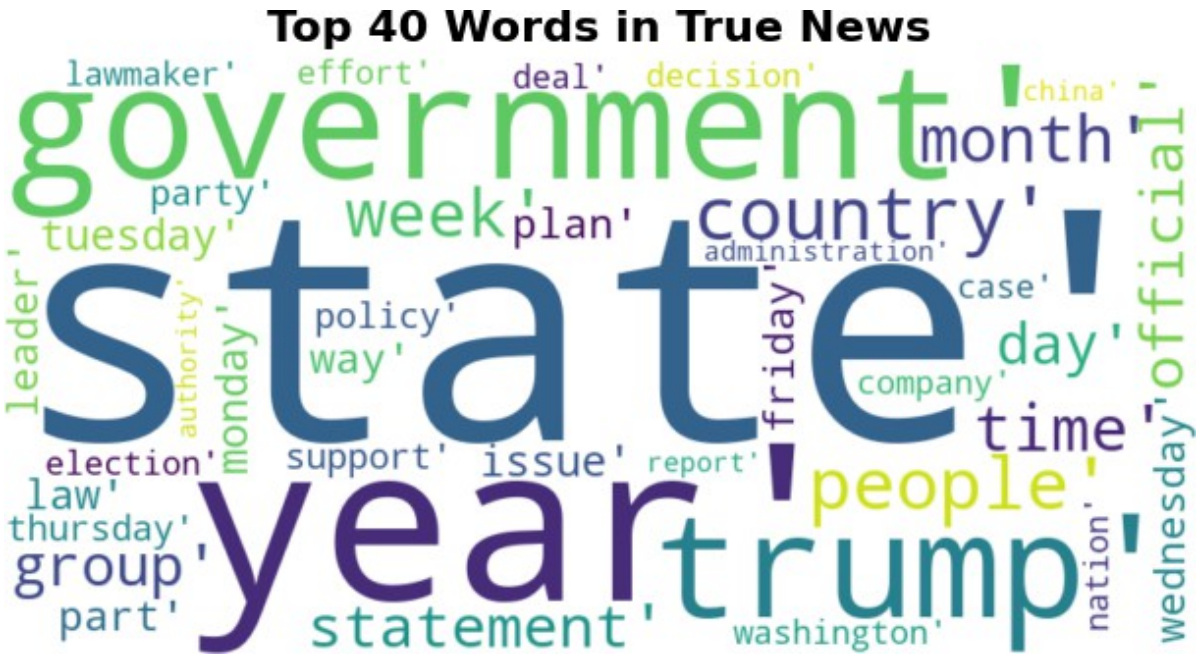


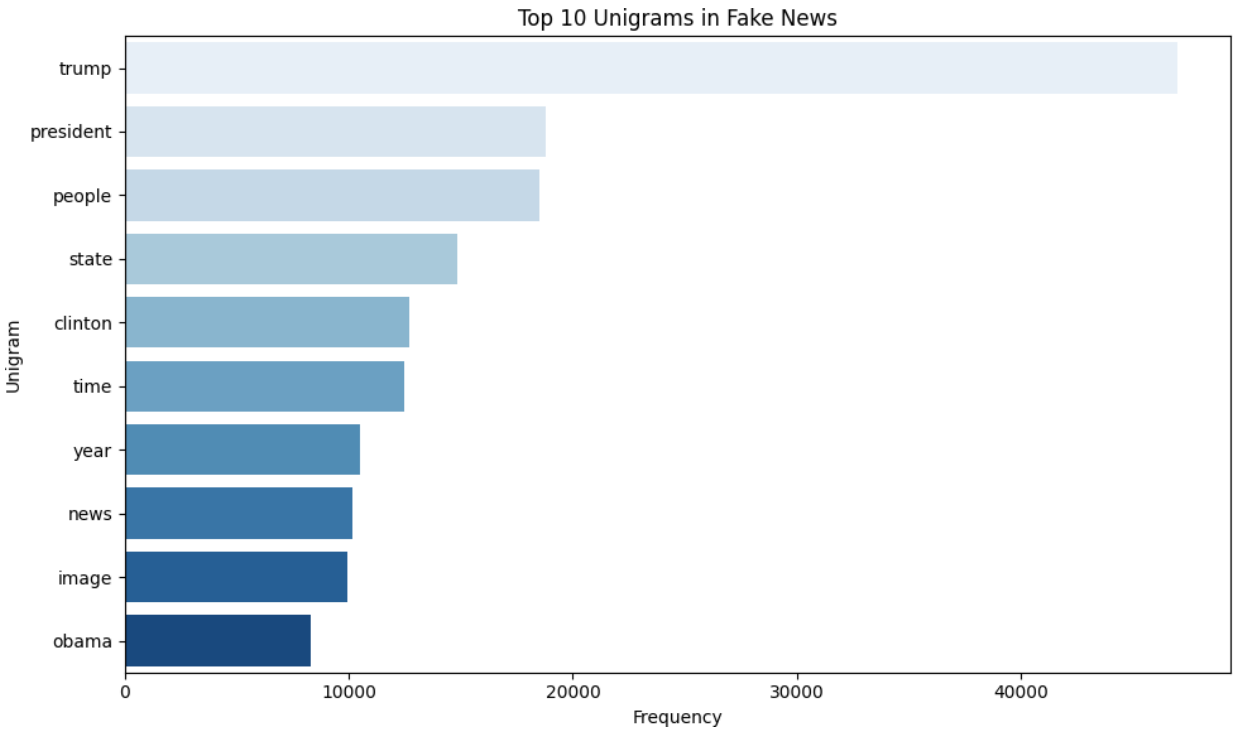

Top 10 Unigrams in True News
trump
state
president
reuters
government
year
house
people
party
election
Unigram
0
5000
10000
15000
20000
25000
30000
Frequency

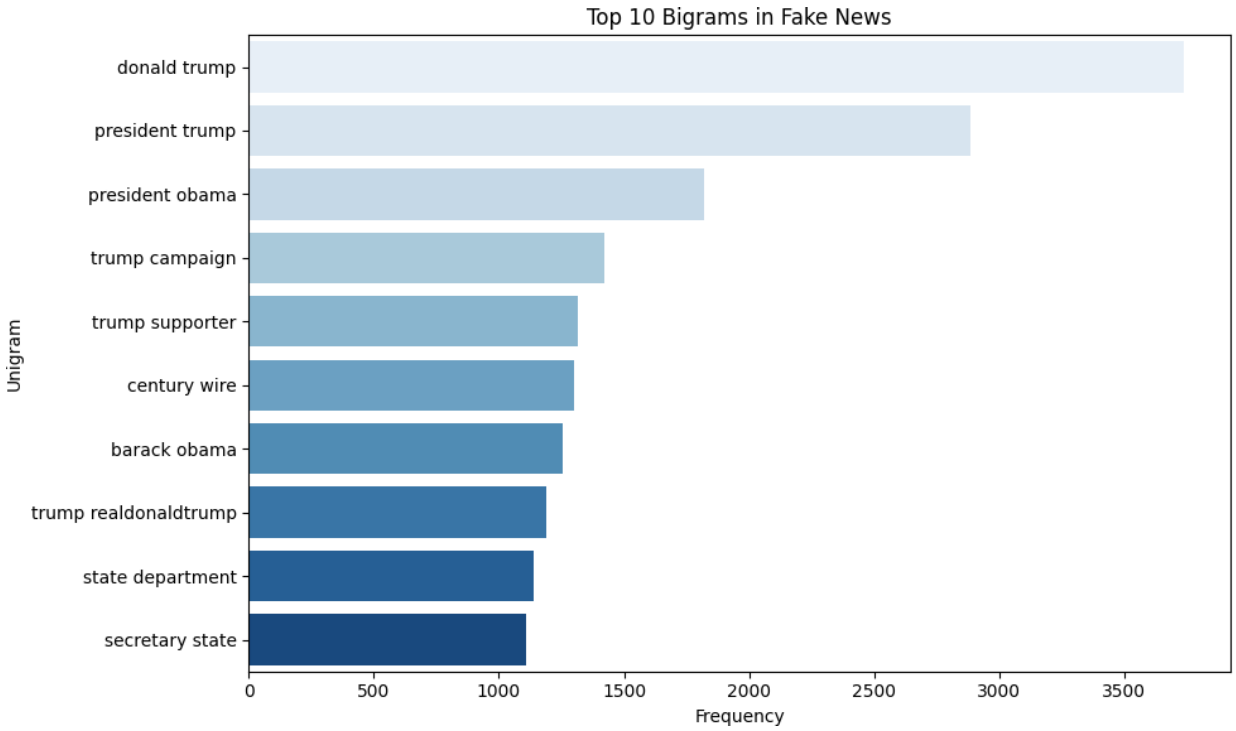

Top 10 Bigrams in True News
donald trump
president trump
washington reuters
president barack
reuters president
president donald
secretary state
house representative
state department
barack obama
Unigram
0
500
1000
1500
2000
Frequency

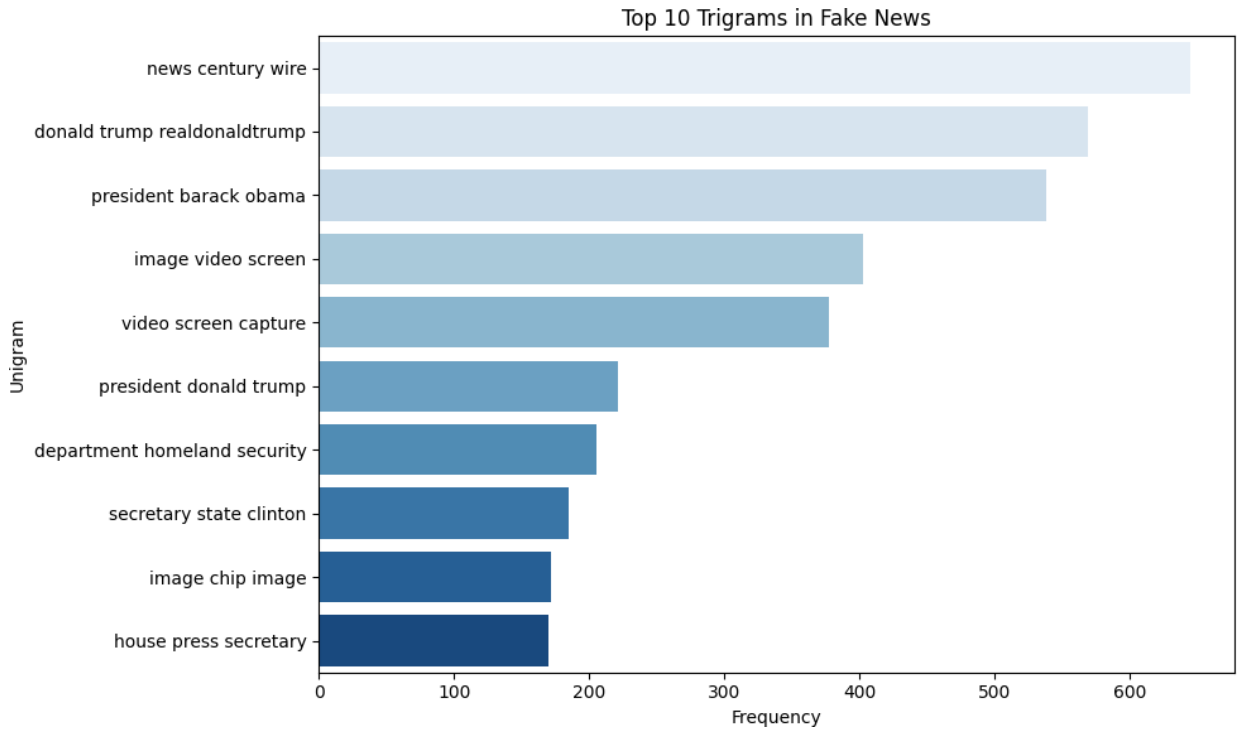


Fake News Linguistic Patterns




Top 10 Unigrams in Fake News
trump
president
people
state
clinton
time
year
news
image
obama
Unigram
0
10000
20000
30000
40000
Frequency


Top 10 Bigrams in Fake News
donald trump
president trump
president obama
trump campaign
trump supporter
century wire
barack obama
trump realdonaldtrump
state department
secretary state
Unigram
0
500
1000
1500
2000
2500
3000
3500
Frequency



Key Pattern Differences

- True News: More institutional references (reuters, washington, government departments)
- Fake News: Higher trump frequency (47K vs 31K), more informal language patterns
- Source Attribution: True news shows formal source citations, fake news shows less structured attribution

CLASSIFICATION MODELING

Performance Metrics Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	93.21%	92.00%	93.93%	92.95%
Decision Tree	84.64%	85.29%	81.93%	83.58%
Random Forest	93.00%	92.00%	94.00%	93.00%

MODEL SELECTION

Best Model: Logistic Regression

Fake News Detection:

- Precision: 94%
- Recall: 93%
- F1-Score: 93%
- Support: 7,045 samples

True News Detection:

- Precision: 92%
- Recall: 94%
- F1-Score: 93%
- Support: 6,425 samples

Primary Reasons for Selection:

- Highest Overall Accuracy: 93.21% outperforms Decision Tree (84.64%) and matches Random Forest
- Best F1-Score: 92.95% demonstrates optimal precision-recall balance
- Superior Precision: 92.00% vs Random Forest's comparable performance
- Computational Efficiency: Faster training and prediction than ensemble methods
- Balanced Class Performance: 93% F1-score for both fake and true news categories
- Interpretability: Clear understanding of semantic feature contributions

CONCLUSION

This semantic classification project successfully demonstrates that Word2Vec embeddings combined with Logistic Regression can achieve exceptional 93.21% accuracy in fake news detection. The comprehensive analysis revealed distinct linguistic patterns: true news emphasizing institutional references (reuters, government, washington) while fake news showed higher frequencies of sensationalized terms and informal language patterns.

The Logistic Regression model's superior performance (93.21% accuracy, 92.95% F1- score) coupled with balanced class detection (93% F1-score for both categories) makes it the optimal choice for production deployment. The semantic approach proved superior to traditional methods by capturing contextual meaning through 300-dimensional vector representations.