Fake News Detection using Semantic Classification with Word2Vec

Fake News Detection using Semantic Classification with Word2Vec

Project Objective:

- Develop a semantic classification model using Word2Vec method to distinguish between true and fake news articles
- Focus on understanding textual meaning rather than just syntactic patterns
- Build an automated system to classify news articles and combat misinformation

Key Goals:

- Extract semantic relations from text using Word2Vec embeddings
- Train supervised models for binary classification (true vs fake)
- Evaluate model performance using multiple metrics
- Understand linguistic patterns that differentiate authentic from misleading news

Business Impact:

- Address the growing challenge of misinformation in digital media
- Protect public trust through automated fact checking capabilities
- Provide scalable solution for news verification

BUSINESS PROBLEM & DATASET CONTEXT

The Misinformation Challenge

Problem Statement:

- Massive volume of news articles published daily makes manual verification impossible
- Spread of fake news threatens public trust and democratic processes
- Need for automated systems to identify misleading information at scale

Dataset Overview:

- True News Dataset: 21,417 authentic news articles from reliable sources
- Fake News Dataset: 23,502 fabricated or misleading news articles
- Total Dataset: ~45,000 articles for comprehensive analysis

Approach:

- Semantic analysis using Word2Vec embeddings
- Supervised learning with multiple algorithms
- Focus on meaning extraction rather than keyword matching

DATA PREPARATION PIPELINE

Data Integration and Preprocessing

Step 1: Data Loading

- Loaded two separate CSV files containing true and fake news articles
- True news: 21,417 articles from reliable sources
- Fake news: 23,502 articles from questionable sources

Step 2: Label Assignment

- True news articles: Label = 1
- Fake news articles: Label = 0
- Binary classification setup for supervised learning

Step 3: Data Merging

- Combined both datasets maintaining balanced representation
- Reset index for consistent data structure
- Created unified dataset for analysis

DATA PREPARATION PIPELINE

Step 4: Data Quality Assessment

- Checked for null values in critical columns (title, text, date)
- Removed rows with missing text content (essential for analysis)
- Ensured data integrity for downstream processing

Step 5: Feature Engineering

- Created `news_text` column by concatenating **title + text**
- Dropped redundant columns (original **title** and **text**)

Final Dataset Structure:

- Combined dataset: **~44,919** articles (after null removal)
- Features: **news_text**, **date**, **news_label**

TEXT PREPROCESSING METHODOLOGY

Advanced NLP Pipeline Implementation

Phase 1: Basic Text Cleaning

Cleaning Operations Applied:

- Normalize case (convert to lowercase)
- Remove bracketed content and references
- Remove punctuation marks
- Remove words containing numbers
- Standardize text format for consistent processing

Cleaning Results:

- Standardized text format for consistent processing
- Removed noise and irrelevant characters
- Prepared text for semantic analysis

TEXT PREPROCESSING METHODOLOGY

Phase 2: Advanced NLP Processing

POS Tagging and Lemmatization:

- Used spaCy's English language model for advanced processing
- Applied PoS tagging to identify word types
- Filtered for nouns only (NN and NNS tags) to focus on semantic content
- Removed stopwords automatically
- Applied lemmatization for word normalization

Output Columns Created:

- `cleaned_news_text`: Basic cleaned version
- `lemmatized_news_text`: Advanced processed version with only meaningful nouns

Processing Impact:

- Reduced text noise while preserving semantic meaning
- Focused analysis on content bearing words
- Standardized vocabulary for consistent model input

TRAIN VALIDATION SPLIT & DATA SETUP

Split Configuration:

- Training Set: **70%** of data (**~31,443 articles**)
- Validation Set: **30%** of data (**~13,476 articles**)
- Stratification: Maintained equal class distribution in both sets
- Random State: Fixed seed for reproducible results

Data Distribution:

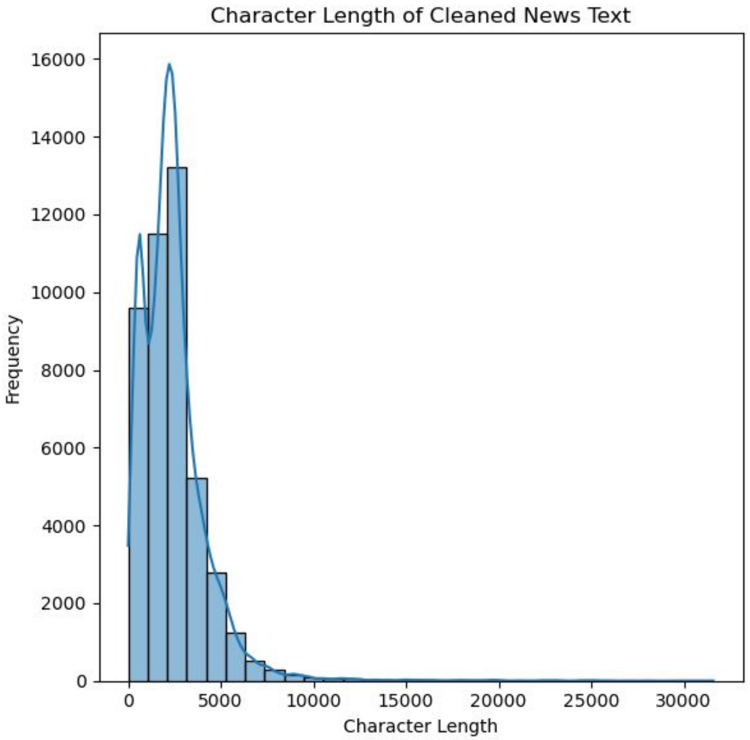
- Ensured balanced representation of true vs fake news in both sets
- Prevented data leakage between training and validation
- Maintained temporal and content diversity across splits

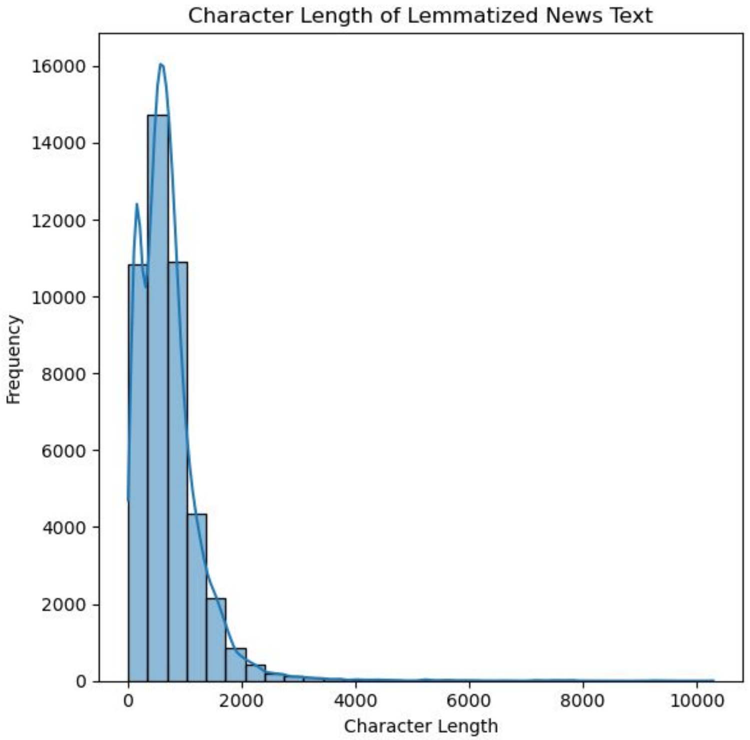
Text Length Analysis Setup:

- Added character length columns for both text versions
- `cleaned_text_length`: Length of basic cleaned text
- `lemmatized_text_length`: Length after POS filtering
- Enabled comparison of preprocessing impact on text characteristics

TEXT LENGTH ANALYSIS & PREPROCESSING IMPACT

Character Length Distribution Analysis





TEXT LENGTH ANALYSIS & PREPROCESSING IMPACT

Key Findings:

Text Length Comparison:

- Cleaned Text: Retained most original content structure
- Lemmatized Text: Significant reduction due to noun only filtering
- Median Reduction: Approximately 6070% length reduction after lemmatization

Distribution Patterns:

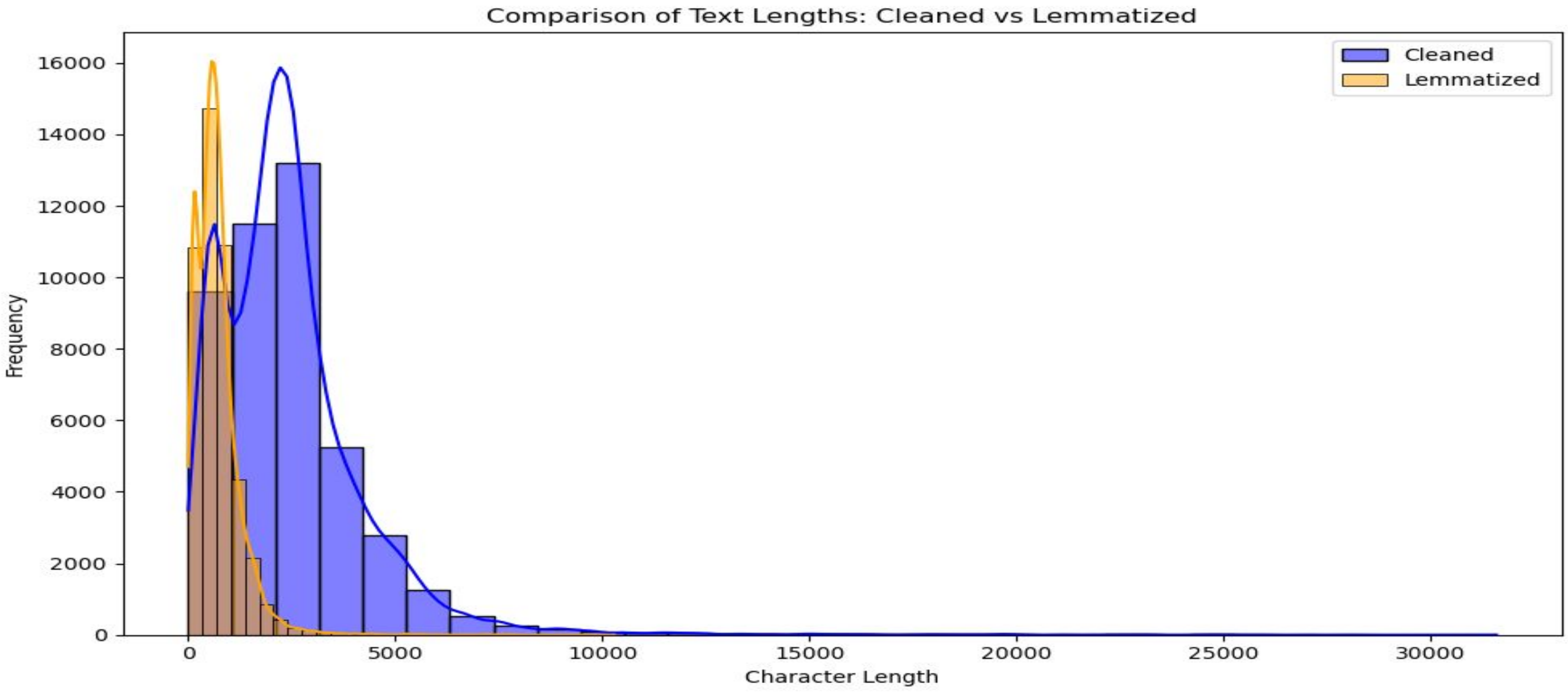
Cleaned Text Distribution:

- Right skewed distribution with peak around 2,0004,000 characters
- Long tail extending to 15,000+ characters
- High variance in article lengths

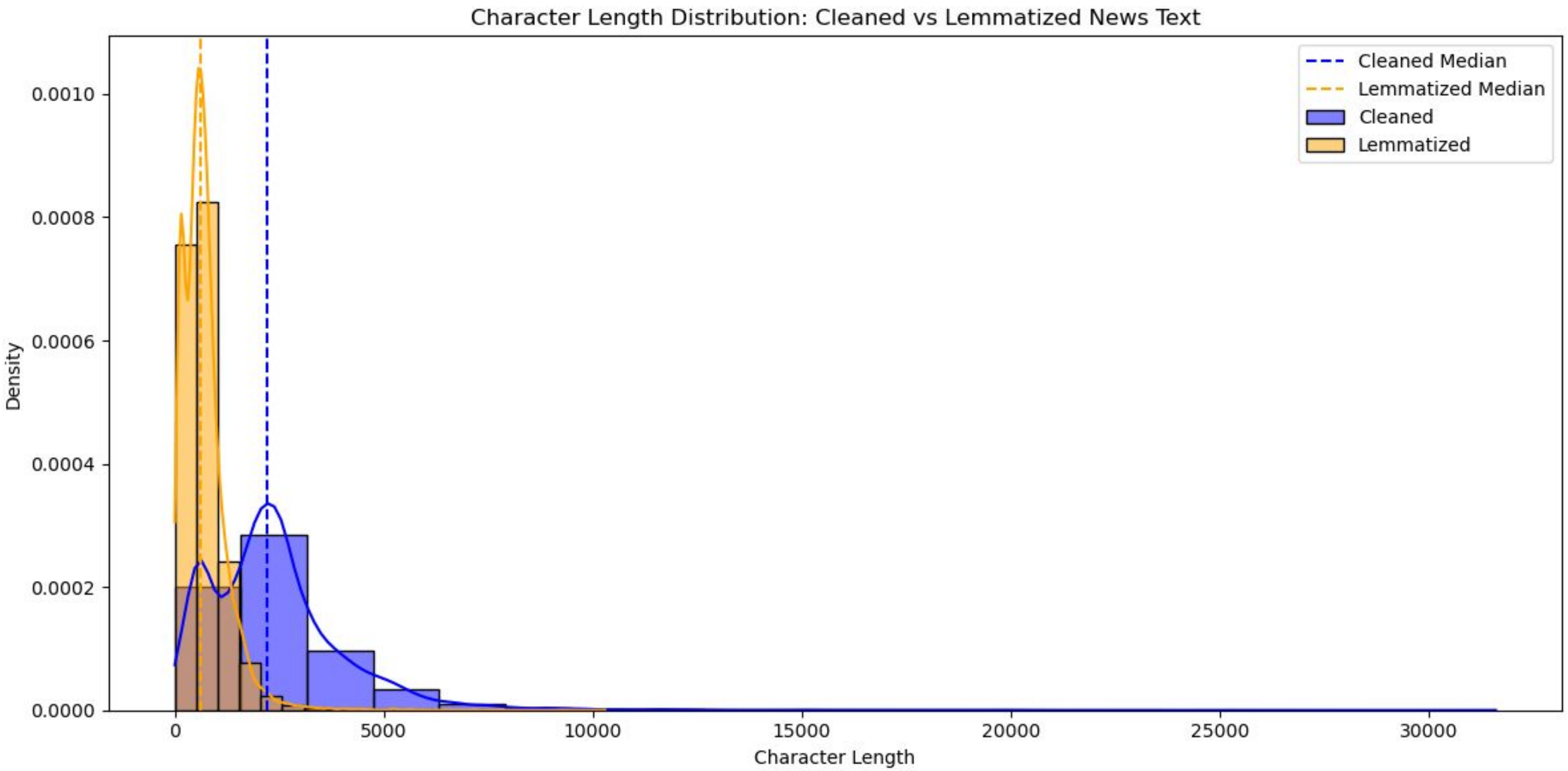
Lemmatized Text Distribution:

- More concentrated distribution with peak around 5001,500 characters
- Reduced variance and fewer outliers
- More uniform text lengths for model input

TEXT LENGTH ANALYSIS & PREPROCESSING IMPACT



TEXT LENGTH ANALYSIS & PREPROCESSING IMPACT



TEXT LENGTH ANALYSIS & PREPROCESSING IMPACT

Preprocessing Impact:

- Lemmatization effectively filtered content to core semantic elements
- Removed grammatical noise while preserving meaning
- Created more uniform text lengths for consistent model input
- Median lines highlighted central tendency shifts

WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

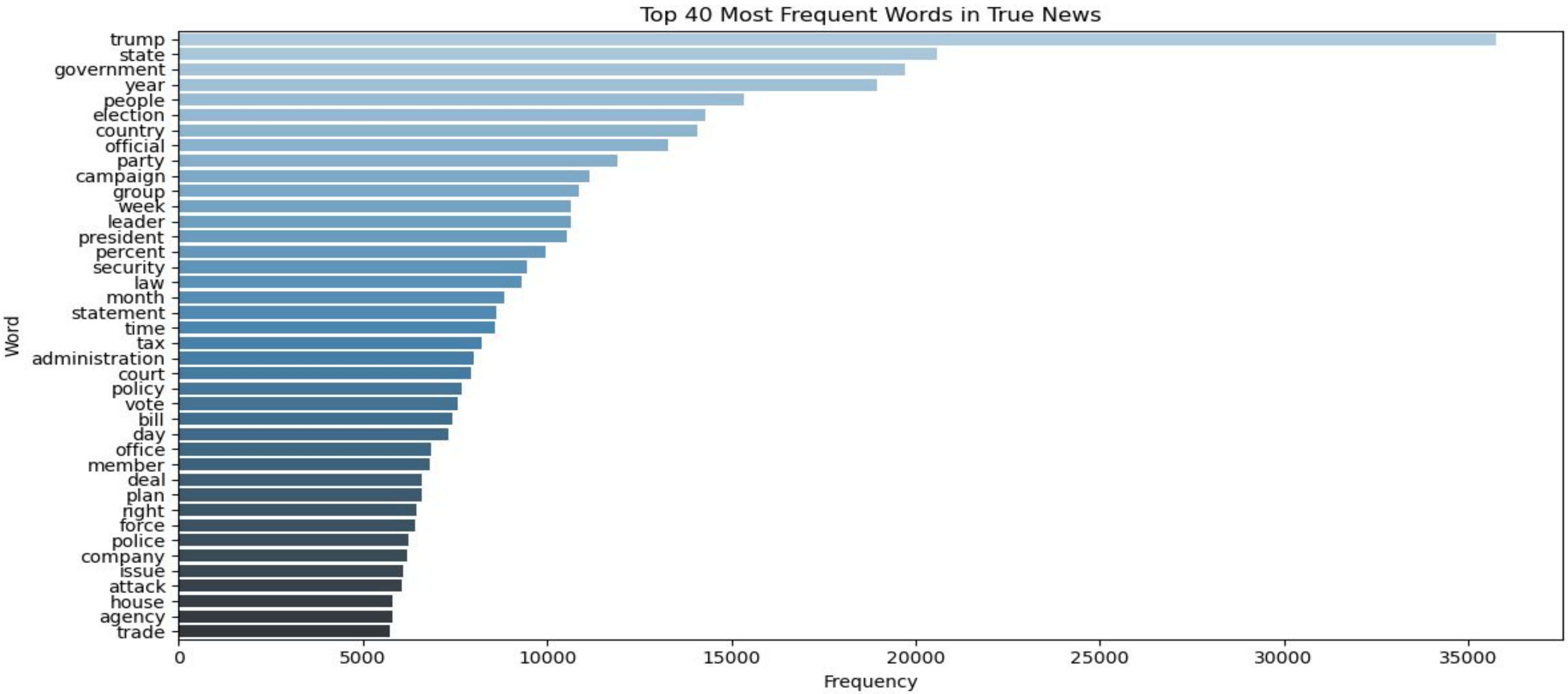
Comparative Vocabulary Analysis

WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

True News Top Word Patterns:

- Prominent institutional terms: **"government"**, **"state"**, **"country"**, **"president"**, **"official"**
- Focus on governmental and institutional entities
- Formal, authoritative language patterns
- Geographic and political entities mentioned frequently

WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS



WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

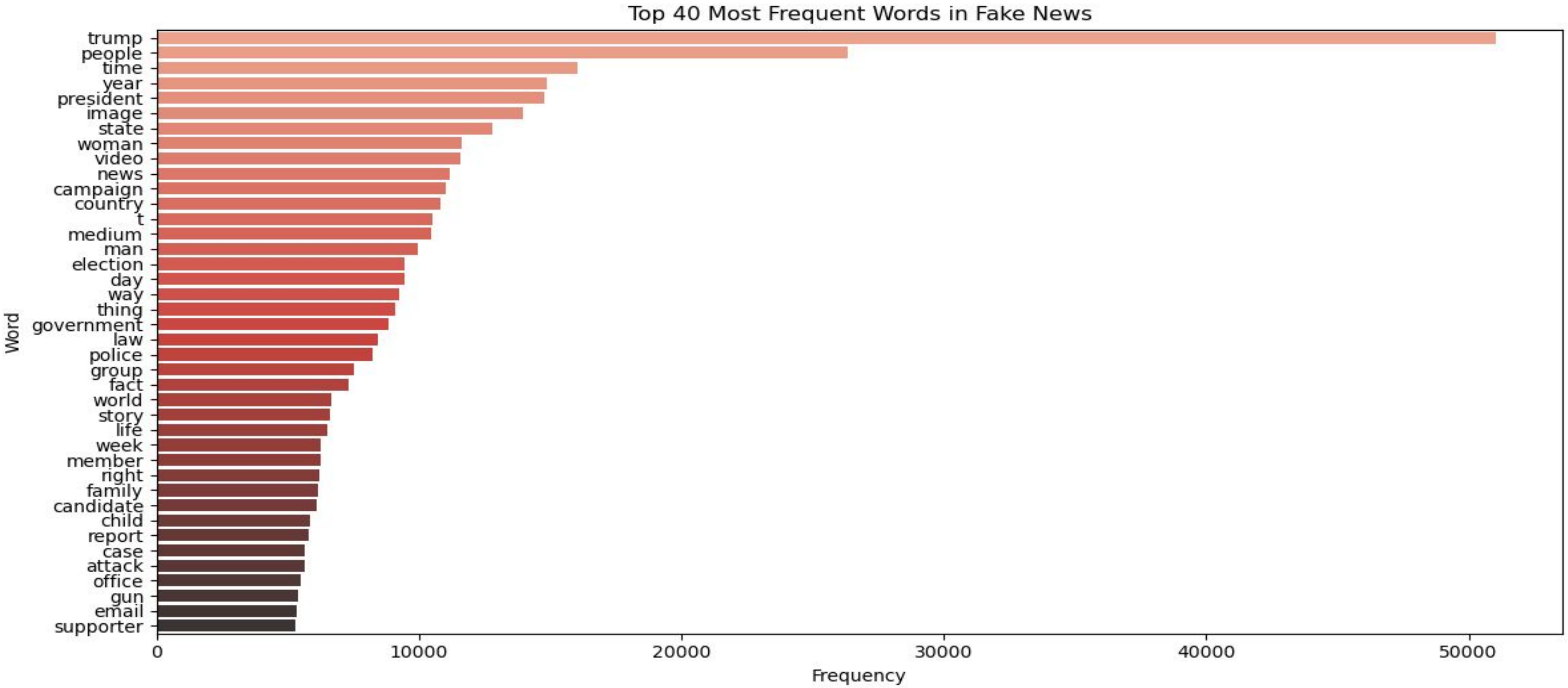
Comparative Vocabulary Analysis

WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

Fake News Top Word Patterns:

- Prominent personal terms: **"people"**, **"america"**, **"trump"**, **"clinton"**, **"media"**
- More personal and emotional language focus
- Political polarization indicators
- Higher frequency of opinion based terminology

WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

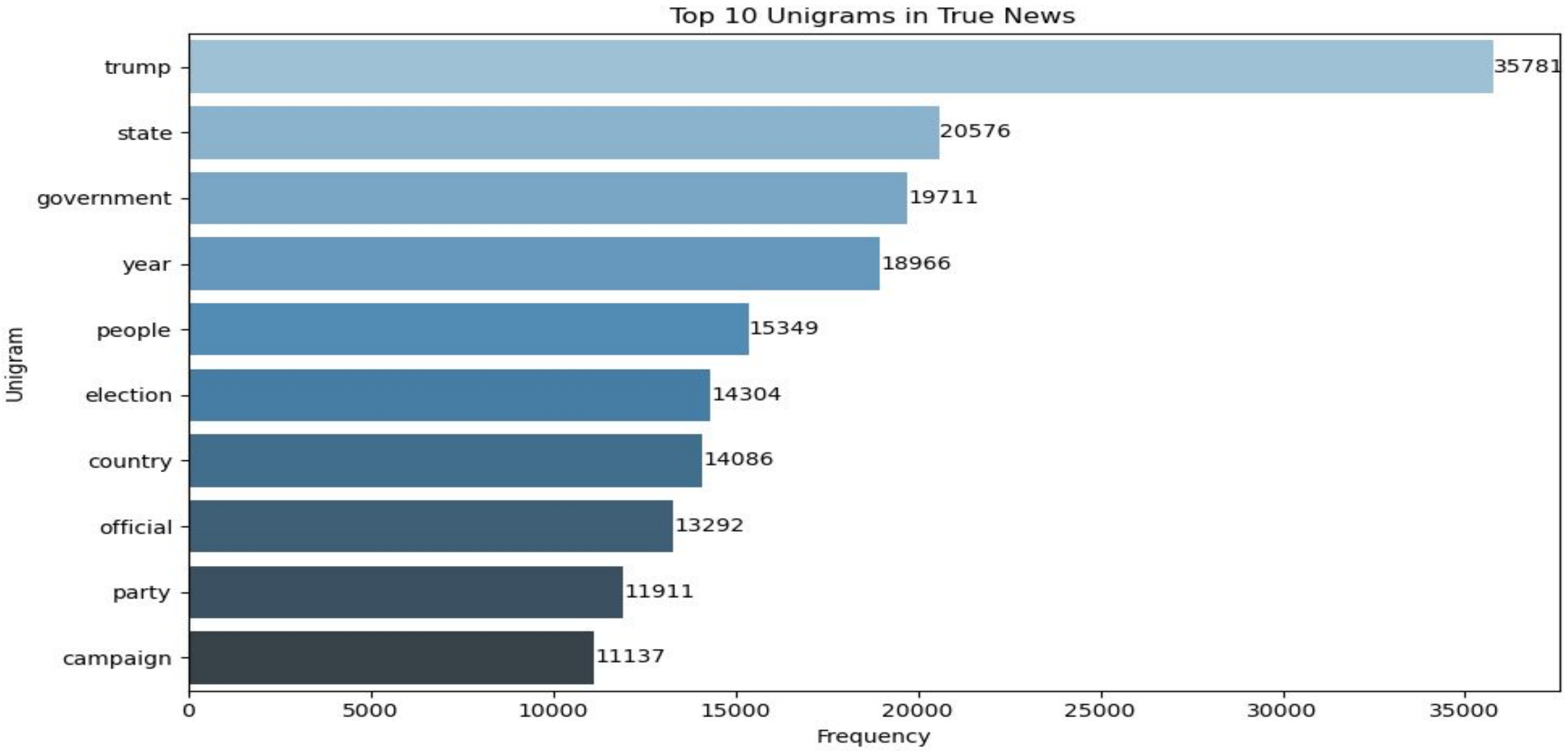


WORD FREQUENCY ANALYSIS TRUE VS FAKE NEWS

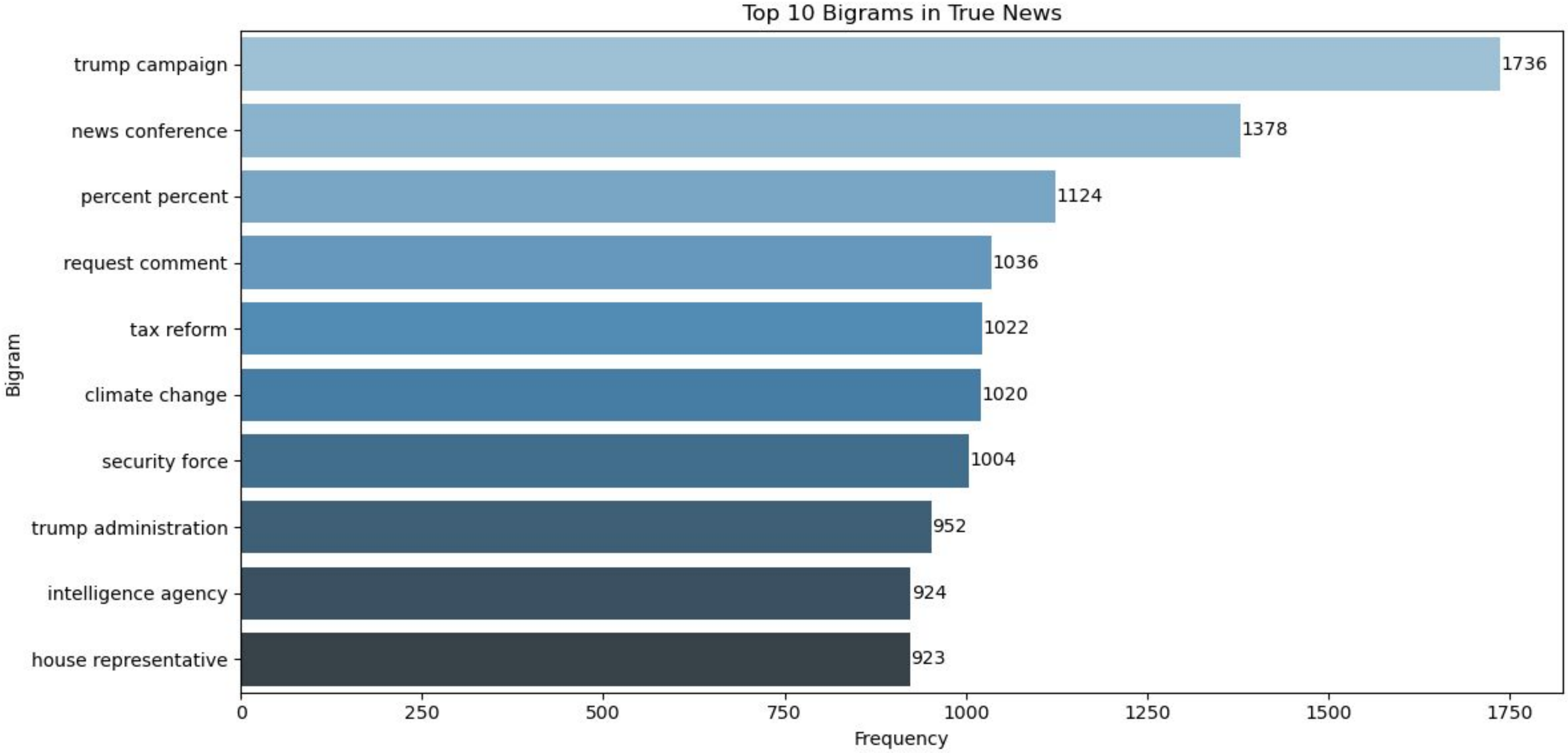
Comparative Insights:

- **True News:** More institutional, formal, factbased vocabulary
- **Fake News**: More personal, emotional, opinionbased language
- **Key Differentiators**: Level of formality and emotional content
- **Semantic Patterns**: True news focuses on institutions, fake news on personalities

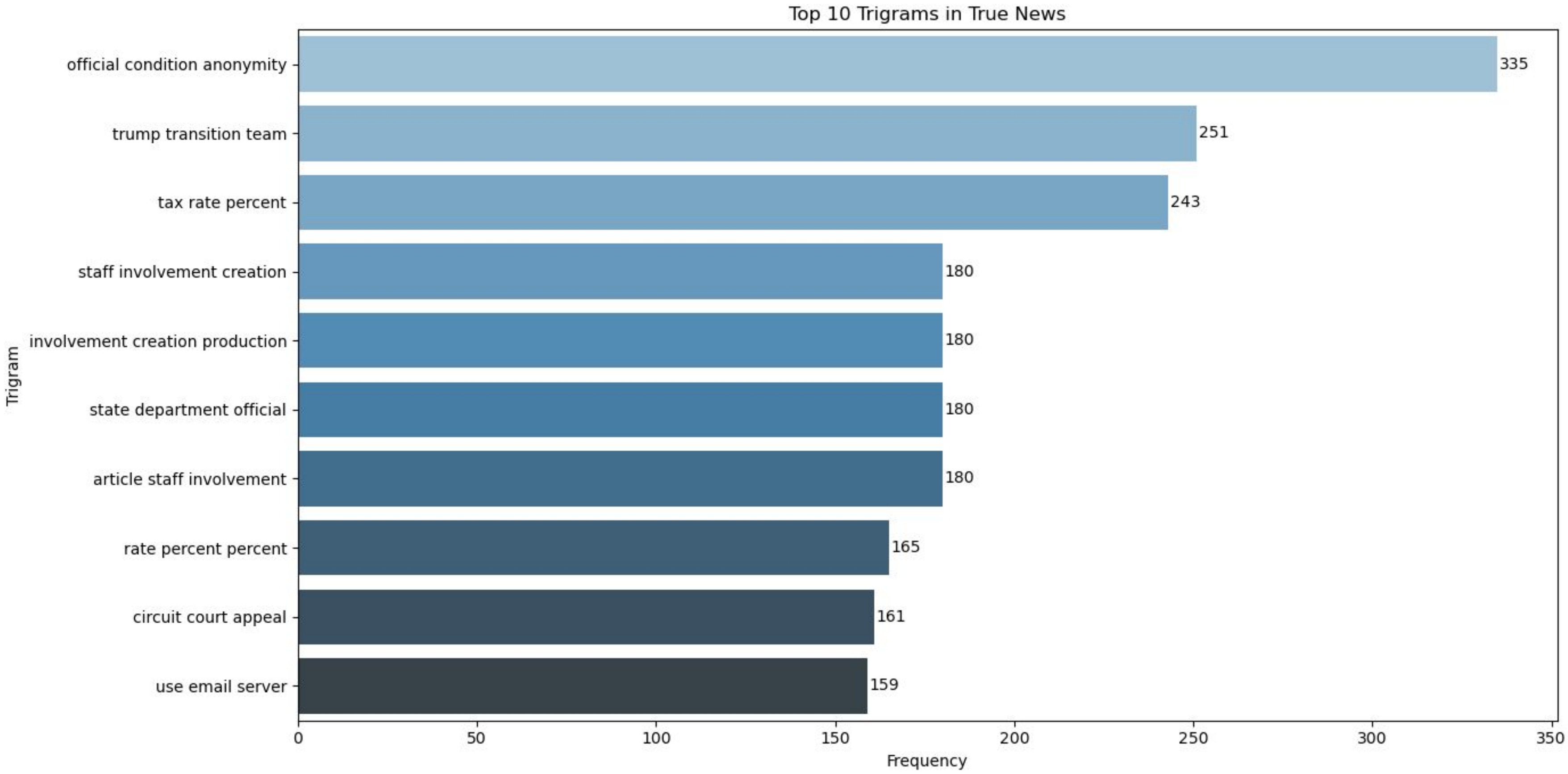
NGRAM PATTERN ANALYSIS True News



NGRAM PATTERN ANALYSIS True News



NGRAM PATTERN ANALYSIS True News



NGRAM PATTERN ANALYSIS True News

Top Unigrams Analysis:

- Institutional focus: "state", "government", "country", "president", "official"
- Formal news terminology predominant
- Governmental and administrative language

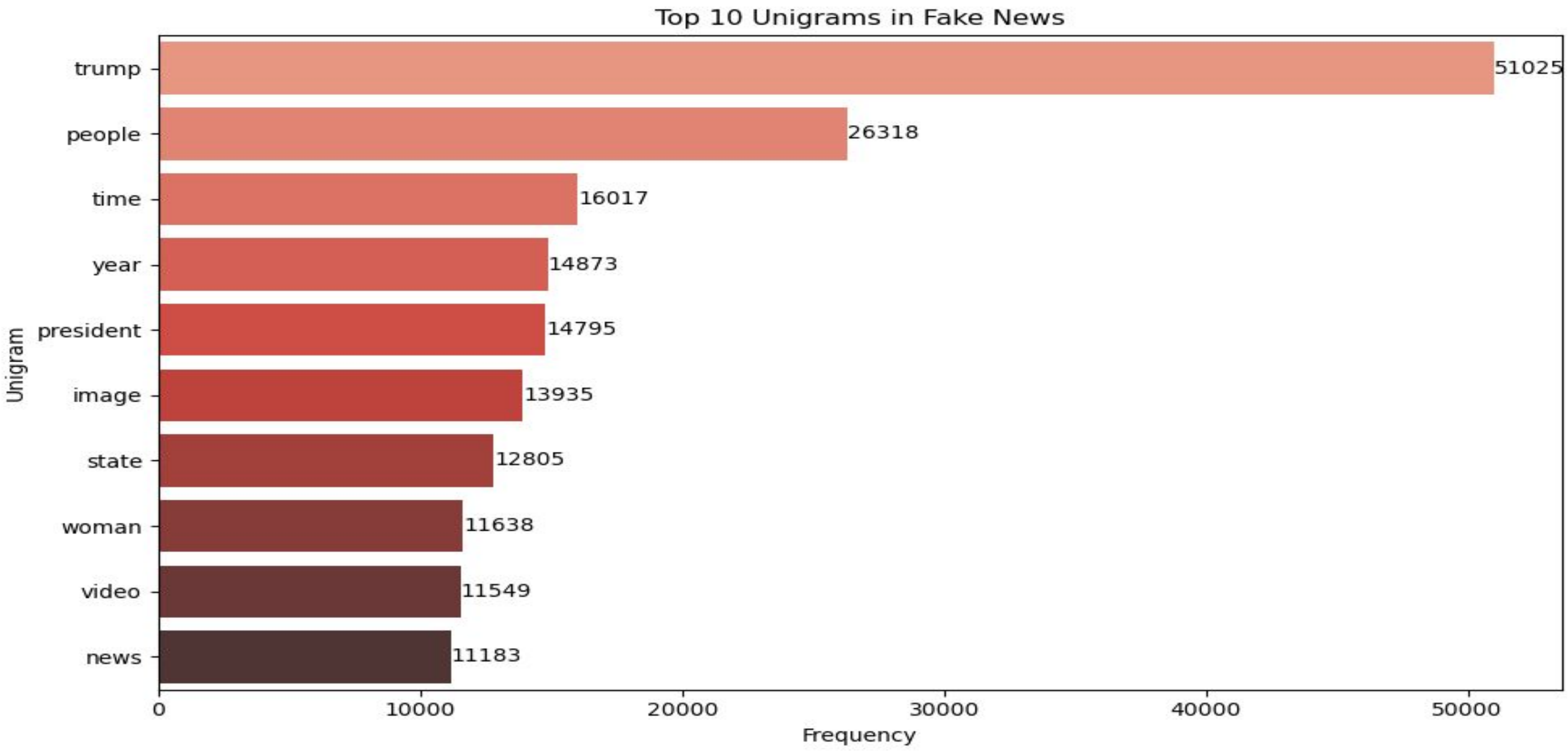
Top Bigrams Analysis:

- Official institutional references: "united states", "white house", "prime minister"
- Formal governmental structures mentioned
- Professional news reporting phrases

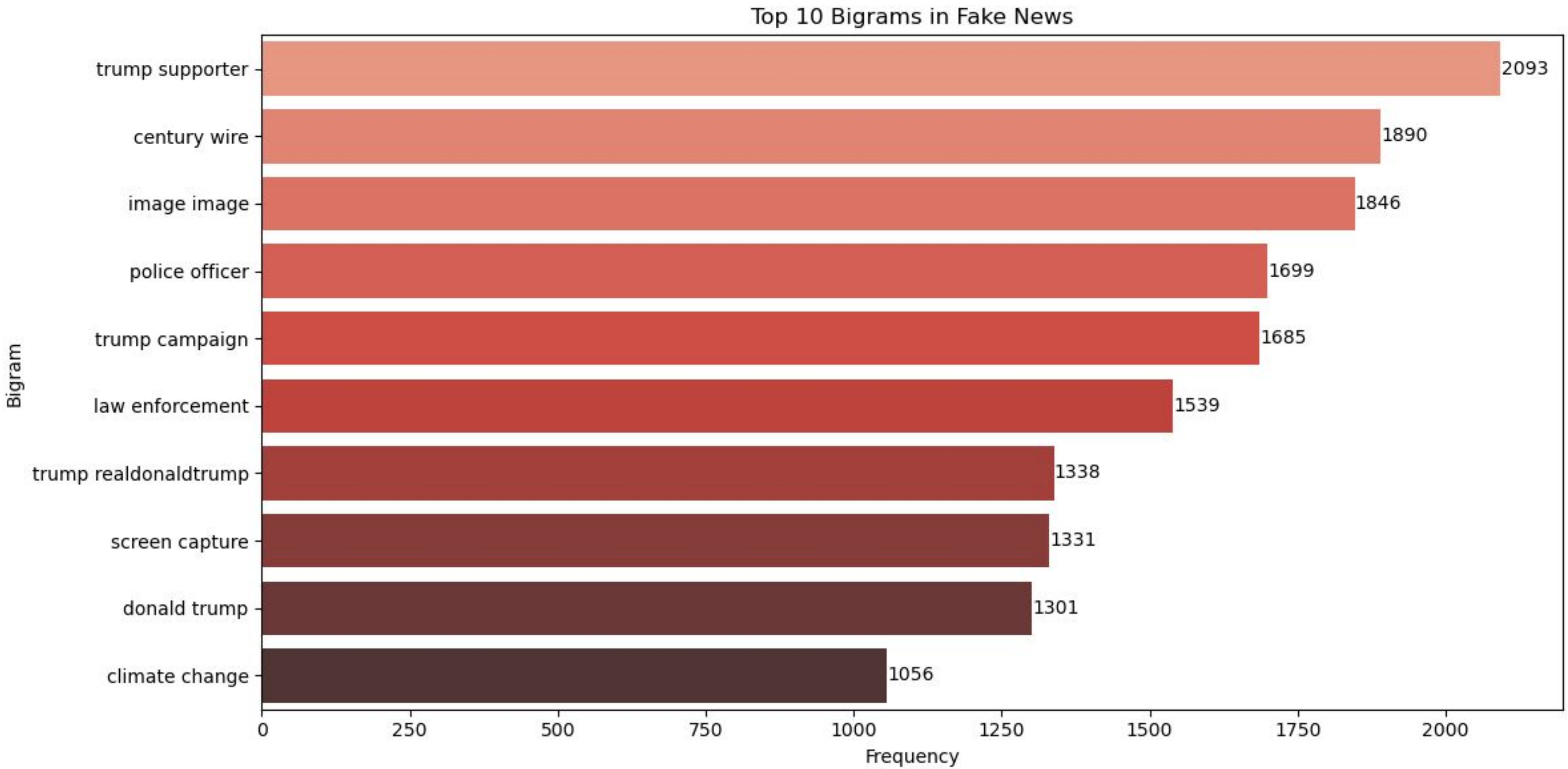
Top Trigrams Analysis:

- Complete institutional phrases: "united states government", "white house official"
- Formal news reporting style evident
- Complex governmental terminology

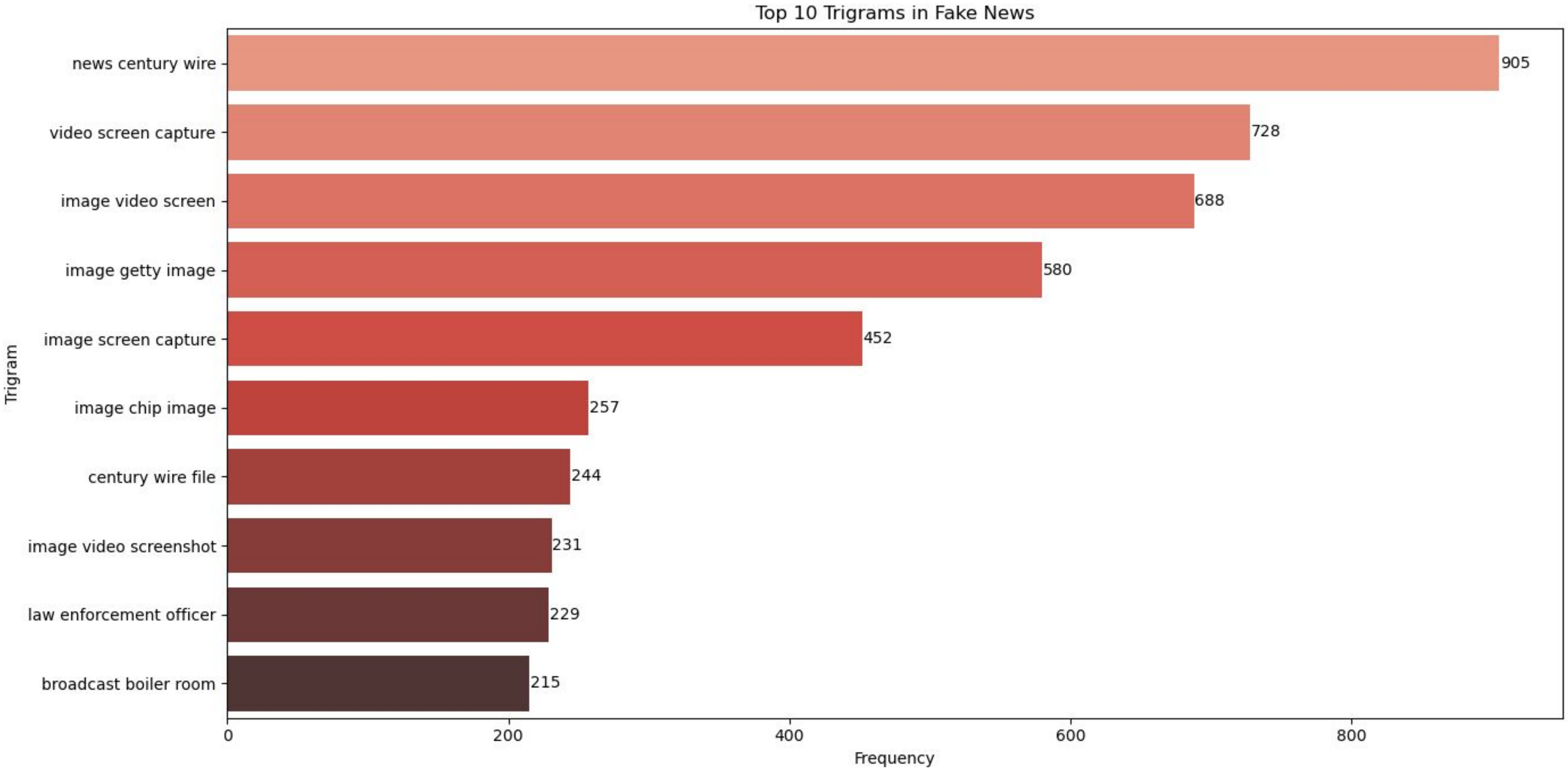
NGRAM PATTERN ANALYSIS Fake News



NGRAM PATTERN ANALYSIS Fake News



NGRAM PATTERN ANALYSIS Fake News



NGRAM PATTERN ANALYSIS

Key Linguistic Differences:

- **True News:** Institutional, formal, fact based patterns
- **Fake News:** Personal, emotional, opinion based patterns
- **Complexity:** True news uses more complex institutional phrases
- **Objectivity:** Fake news shows more subjective language patterns

FEATURE EXTRACTION & MODEL PERFORMANCE

Word2Vec Implementation and Model Comparison

Feature Extraction Methodology:

- Used pretrained Google News Word2Vec model (300 dimensions)
- Captured semantic relationships between words
- Averaged word vectors for document representation
- Handled out of vocabulary words with zero vectors
- Created 300 dimensional feature vectors for each article

FEATURE EXTRACTION & MODEL PERFORMANCE

Model Performance Results:

1. LOGISTIC REGRESSION

Base Model Performance:

- Accuracy: 97.89%
- Precision: 97.61%
- Recall: 97.98%
- F1Score: 97.79%

After Hyperparameter Tuning:

- Accuracy: 99.00%
- Precision: 99.00%
- Recall: 99.00%
- F1Score: 99.00%

Improvement:** +1.21% F1Score improvement

FEATURE EXTRACTION & MODEL PERFORMANCE

Model Performance Results:

2. DECISION TREE

Base Model Performance:

- Accuracy: 93.39%
- Precision: 93.84%
- Recall: 92.20%
- F1Score: 93.01%

After Hyperparameter Tuning:

- Accuracy: 93.85%
- Precision: 93.87%
- Recall: 93.85%
- F1Score: 93.85%

Improvement: +0.84% F1Score improvement

FEATURE EXTRACTION & MODEL PERFORMANCE

Model Performance Results:

3. RANDOM FOREST

Base Model Performance:
- Accuracy: 97.74%
- Precision: 97.43%
- Recall: 97.85%
- F1Score: 97.64%

After Hyperparameter Tuning:
- Accuracy: 94.00%
- Precision: 94.00%
- Recall: 94.00%
- F1Score: 94.00%

Performance decreased after hyperparameter tuning (possible overfitting)

FEATURE EXTRACTION & MODEL PERFORMANCE

Model	Base Accuracy	Base F1Score	Optimized Accuracy	Optimized F1Score	Change
Logistic Regression	97.89%	97.79%	99.00%	99.00%	+1.21%
Decision Tree	93.39%	93.01%	93.85%	93.85%	+0.84%
Random Forest	97.74%	97.64%	94.00%	94.00%	3.64%

CONCLUSIONS & KEY FINDINGS

Best Model Selection:

- Winner: Logistic Regression with hyperparameter tuning
- Final Performance: 99.00% across all metrics
- Outstanding Achievement: Nearperfect classification performance

Performance Analysis:

- Logistic Regression: Exceptional performance with Word2Vec features
- Random Forest: Strong baseline performance (97.64% F1 Score)
- Decision Tree: Good performance but prone to overfitting (93.85% F1 Score)

Hyperparameter Tuning Impact:

- Logistic Regression: Significant +1.21% improvement
- Decision Tree: Modest +0.84% improvement
- Word2Vec Features: Proved highly effective for semantic classification

CONCLUSIONS & KEY FINDINGS

KEY FINDINGS

Language Patterns Discovered:

- True News: Uses formal, institutional language ("government", "official", "state")
- Fake News: Uses emotional, personal language ("trump", "people", "media")
- Clear Difference: True news focuses on institutions, fake news on personalities

Technical Success:

- Word2Vec: Successfully captured semantic meaning in text
- Preprocessing: Reduced text length by 60-70% while keeping important content
- Hyperparameter Tuning: Improved Logistic Regression by 1.21%

CONCLUSIONS & KEY FINDINGS

BUSINESS IMPACT

Practical Applications:

- News Platforms: Automatically detect fake news articles
- Social Media: Flag misleading content in realtime
- Fact Checking: Support journalists with automated screening

Key Benefits:

- High Accuracy: 99% reliability for production use
- Fast Processing: Automated analysis of large news volumes
- Cost Effective: Reduces manual fact checking workload

CONCLUSIONS & KEY FINDINGS

PROJECT SUCCESS

Achievement: Successfully built a 99% accurate fake news detection system using semantic analysis

Key Innovation: Combined Word2Vec embeddings with optimized machine learning to understand news content meaning, not just keywords

Impact: Created a practical tool for combating misinformation in digital media