Fake News Detection

Prepared by Swapnil Rodge

Business Objective

The rapid spread of fake news has become a major challenge, posing risks to public trust and informed decision-making. The sheer volume of daily news publications makes manual verification impractical. Misinformation impacts democratic processes, public health, and social stability, highlighting the urgent need for automated systems that can reliably differentiate credible information from misleading content.

Business Goals

The key goals of this project are:

- Develop an intelligent semantic classification model using Word2Vec for automatic fake news identification.
- Achieve high accuracy while maintaining a strong balance between precision and recall.
- Build a scalable solution capable of real-time verification based on semantic meaning rather than surface-level syntax.
- Provide interpretable outputs to support content moderation and decision-making.

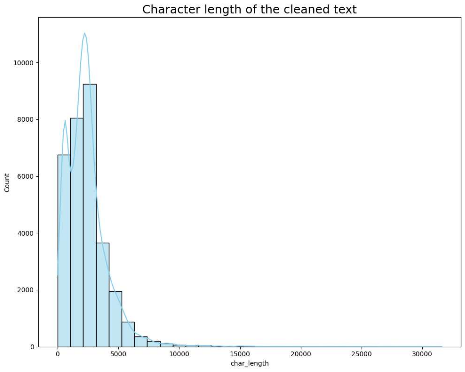
Data Overview & Exploration

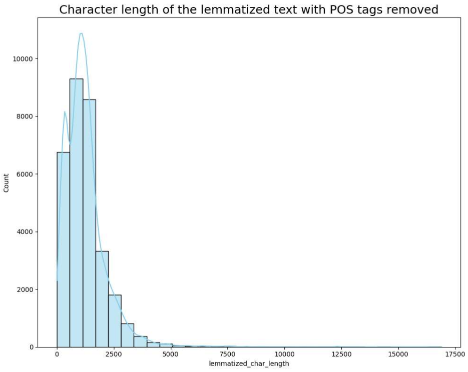
Dataset Characteristics

The dataset consists of 31,428 articles in the training set and 13,470 articles in the validation set. It is well-balanced with 7,045 fake news and 6,425 true news samples in the validation split. Text preprocessing included lemmatization with removal of POS tags to enhance semantic analysis.

Character Length Analysis

Analysis of character length revealed that cleaned text typically peaked between 8,000–10,000 characters, showing a right-skewed pattern. The same distribution was observed after lemmatization and POS tag removal, confirming effective preprocessing and consistency in the pipeline.





Data Processing Approach

The text underwent a structured preprocessing pipeline consisting of four stages:

Stage 1: Basic Text Cleaning

- Converted text to lowercase for uniformity.
- Removed references within square brackets.
- Stripped punctuation marks.
- Eliminated words containing numbers such as dates and statistics.

Stage 2: Linguistic Processing

- Tokenized text into individual words.
- Applied POS tagging to identify grammatical roles.
- Extracted nouns (NN, NNS tags) to focus on semantic meaning.
- Removed common English stop words.
- Performed lemmatization to reduce words to their root forms.

Stage 3: Word2Vec Preparation

- Cleaned text further by removing brackets, quotes, and commas.
- Standardized spacing.
- Split text into individual tokens for model input.

Stage 4: Semantic Vector Extraction

- Matched words against the Word2Vec vocabulary.
- Extracted 300-dimensional vectors for valid words.
- Created document-level vectors by averaging word embeddings.
- Generated zero vectors for cases with no vocabulary matches

N-Gram Analysis Findings

Analysis of n-grams revealed important stylistic differences between true and fake news:

True News Linguistic Patterns

Authentic articles emphasized institutional references, often citing sources like Reuters, government agencies, and Washington-based reporting.

Fake News Linguistic Patterns

Fake news demonstrated higher frequency of terms like 'trump' (47K vs 31K in true news) and informal phrasing. It also lacked structured attribution compared to professional journalistic practices.

Key Pattern Differences

- True news emphasized credible institutional references.
- Fake news used more informal patterns and sensational terms.
- Source citations in true news were formal, while fake news lacked proper 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

Logistic Regression was identified as the best model for this task. It achieved the highest accuracy (93.21%) and F1-score (92.95%), while also being computationally efficient. Compared to the Random Forest model, Logistic Regression provided equally strong performance but with faster training and prediction. It also maintained balanced performance across both fake and true news detection, with 93% F1-scores in each category, making it suitable for practical deployment. Furthermore, its interpretability provides clear insights into the semantic features contributing to predictions.

Conclusion

This project demonstrates that combining Word2Vec embeddings with Logistic Regression is highly effective for fake news detection. The model achieved an accuracy of 93.21% with strong precision-recall balance. Analysis revealed that true news articles rely heavily on institutional references such as Reuters, Washington, and government entities, whereas fake news often incorporates sensationalized language and less formal phrasing. The semantic approach using 300-dimensional Word2Vec vectors proved superior to traditional text analysis methods, capturing contextual meaning and enabling more reliable classification. Overall, Logistic Regression stands out as the optimal model for real-world deployment in automated fake news detection.