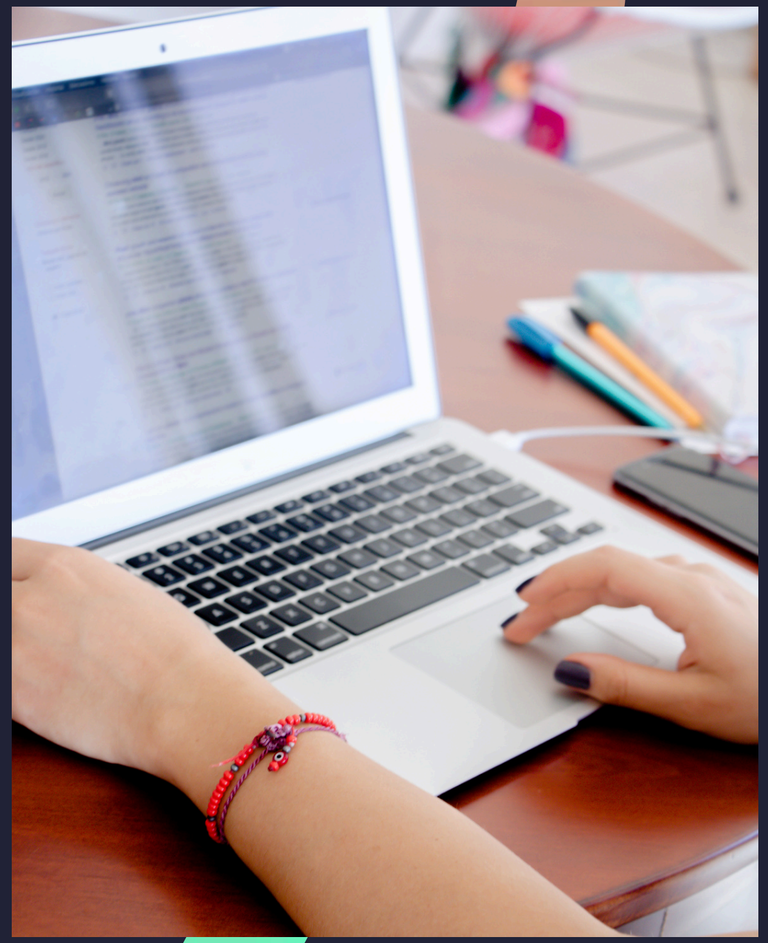
FAKE NEWS DETECTOR

Done by:
Achuth Akilesh



Introduction

- Objective: The project aims to detect whether a given news article is real or fake using natural language processing (NLP) and deep learning.
- Approach: It preprocesses news text data, transforms it using GloVe embeddings, and classifies it using an LSTM-based neural network.
- User Flow: Users can upload a dataset through the Streamlit UI, the model preprocesses and classifies the articles, and displays prediction results along with accuracy.

PROBLEM STATEMENT.

To develop an automated system that accurately detects fake news articles using NLP and deep learning techniques.

ABOUT THE PROJECT.

Overview

This project aims to develop an advanced fake news detection system leveraging deep learning techniques, specifically Long Short-Term Memory (LSTM) networks and Bidirectional Encoder Representations from Transformers (BERT). By analyzing textual data, the system identifies patterns and contextual cues to distinguish between credible and misleading news articles, addressing the growing challenge of misinformation in digital media.

Methodology

- Data Collection: Curate a diverse dataset of labeled news articles (real and fake) from reliable sources.
- Preprocessing: Clean and tokenize text, remove noise (e.g., stopwords, punctuation), and prepare data for model input.
- Model Architecture:
  - LSTM: Processes sequential text data, capturing temporal dependencies and long-term patterns in news content.
  - BERT: Extracts contextual embeddings, understanding nuanced semantics and relationships within text.
- Training and Evaluation: Train models on labeled data, optimize hyperparameters, and evaluate performance using metrics like accuracy, precision, recall, and F1-score.
- Comparison: Analyze the strengths of LSTM (sequential modeling) and BERT (contextual understanding) for fake news detection.

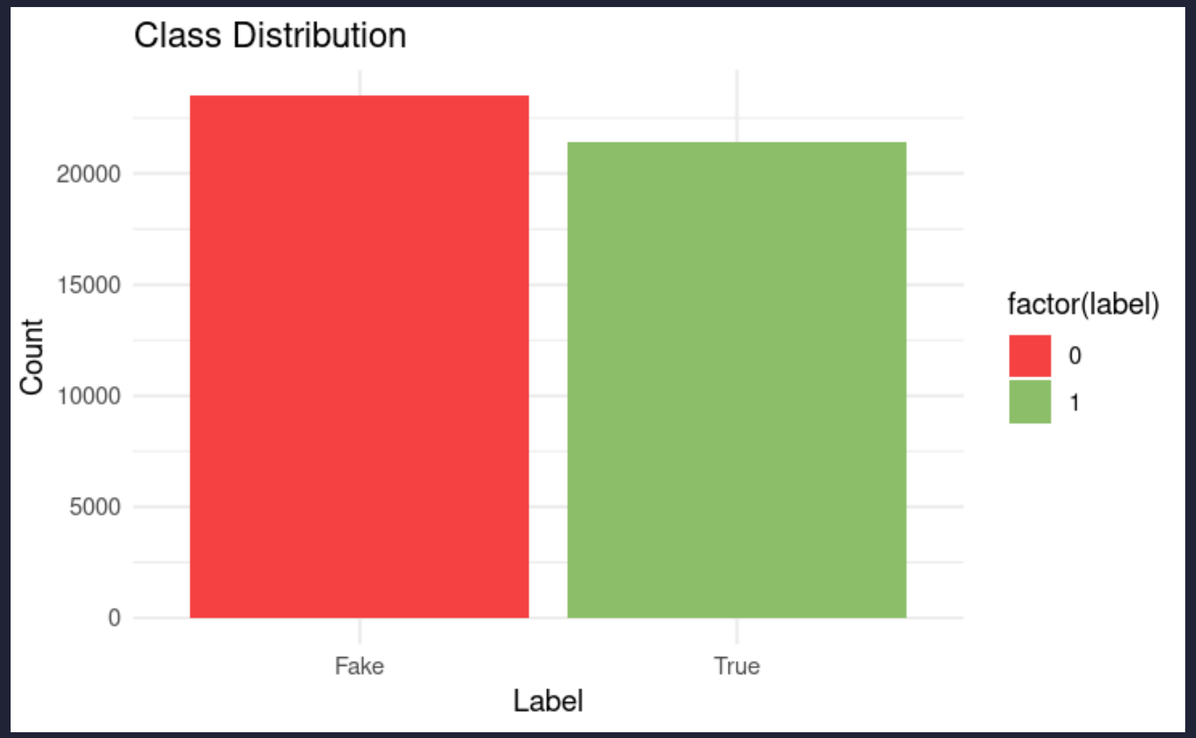
Research Goals

To preprocess and clean textual news data for effective feature extraction and model input.

To explore and integrate word embeddings (GloVe) for capturing contextual meaning of words in news articles.

ANALYSIS HIGHLIGHTS

Class Distribution



01 Dataset shows a slight class imbalance, with more fake news samples than real.

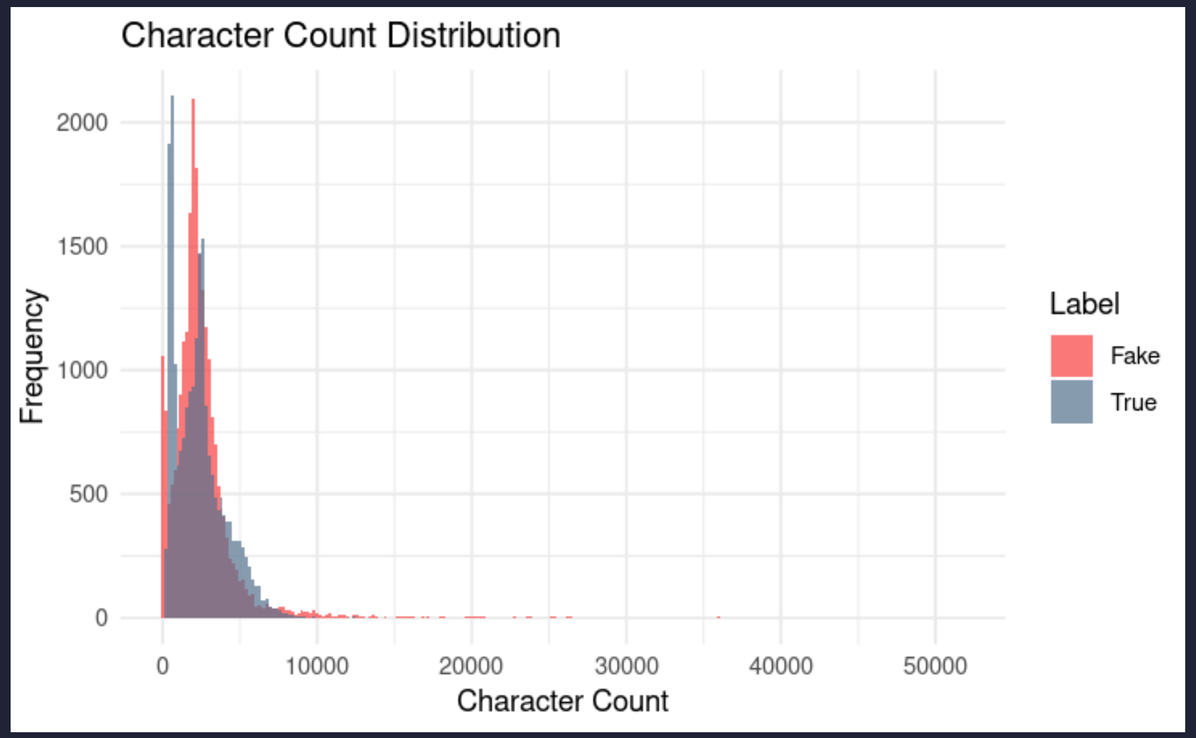
02 Important for model training – class balancing techniques like resampling or class weighting might be necessary.

03 Balanced or near-balanced datasets help in preventing biased models.

04 The visual indicates the need for equal representation in training/testing splits.

05 Reinforces the widespread prevalence of misinformation.

Character Count Distribution



01 Fake news articles tend to have shorter character lengths compared to true news.

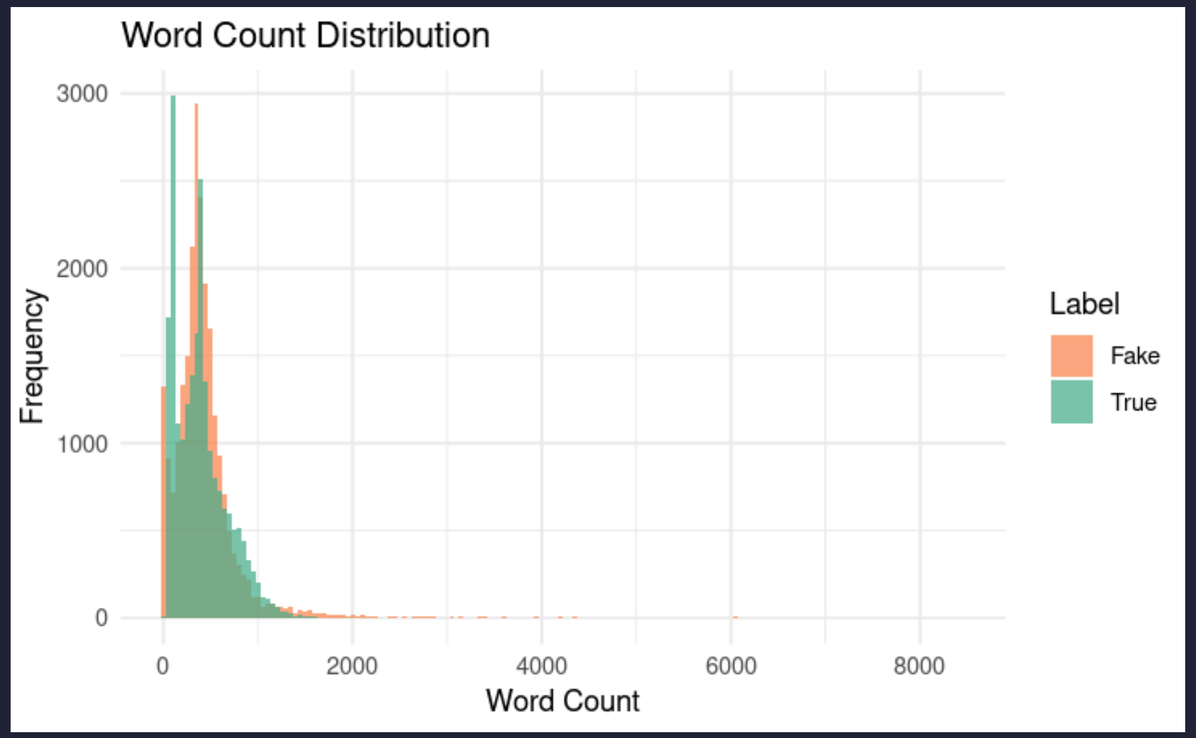
02 High-density region below 5000 characters — useful for feature engineering.

03 The long tail shows a few very lengthy articles, potentially anomalies or special cases.

04 Can use character count as a simple yet powerful feature in predictive modeling.

05 Indicates less depth or content richness in fake news.

Word Count Distribution



01 True news articles generally contain more words than fake ones, suggesting more elaboration.

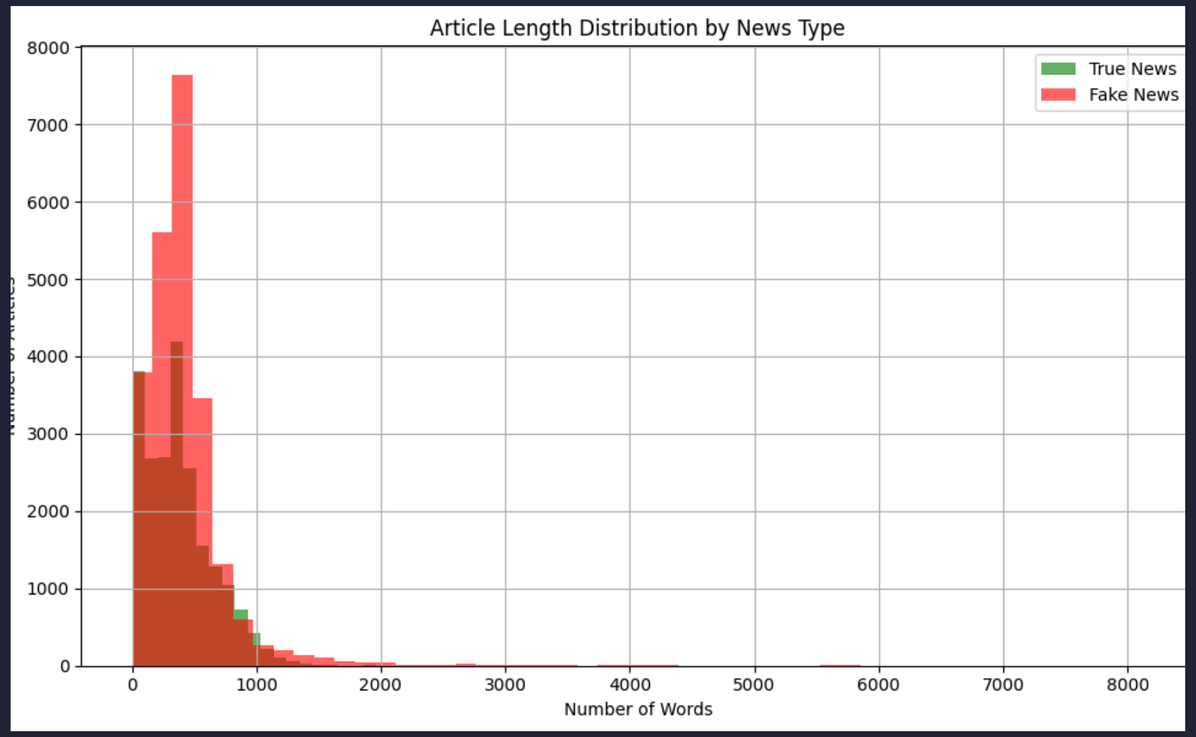
02 Peaks are mostly under 1000 words, with fake news peaking earlier in distribution.

03 Reinforces that fake articles are short and possibly sensational.

04 A clear density difference provides a meaningful signal for ML classifiers.

05 Combining word count with character count can improve text complexity analysis.

Article Length Distribution



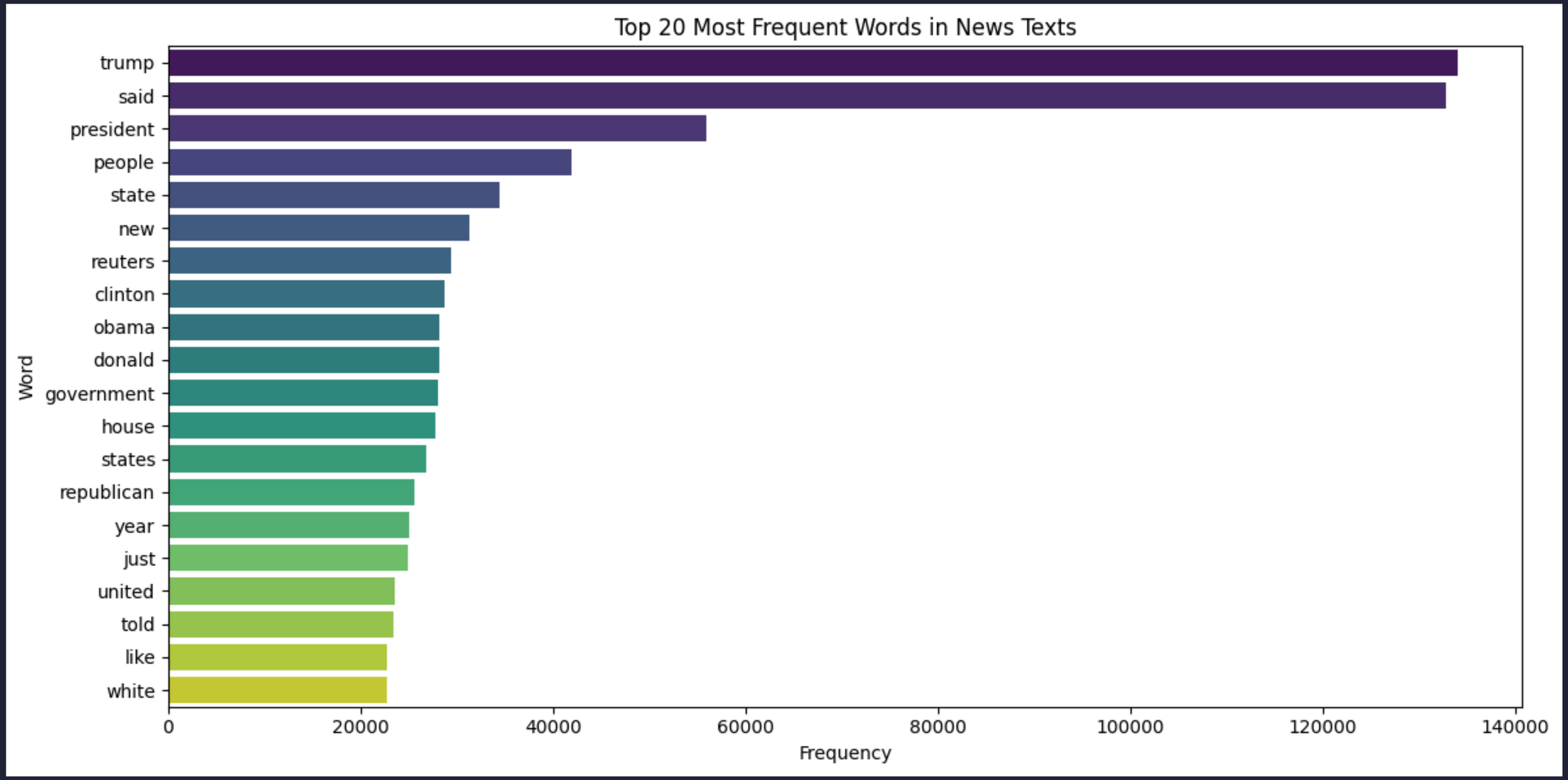
01 Fake news articles dominate the lower word count bins, especially under 1000 words.

02 True news articles show a more spread-out distribution — greater variance in length.

03 Indicates deeper reporting and analysis in genuine content.

04 This difference is an essential feature to distinguish news quality.

05 Useful for data preprocessing: setting minimum length thresholds for filtering low-value content.


Top 20 Most Frequent Words in News Texts
trump
said
president
people
state
new
reuters
clinton
obama
donald
government
house
states
republican
year
just
united
told
like
white
Word
0
20000
40000
60000
80000
100000
120000
140000
Frequency

Top 20 Frequent Words

01 The most common words are "trump", "said", "president", "people", "state" – politics-heavy context.

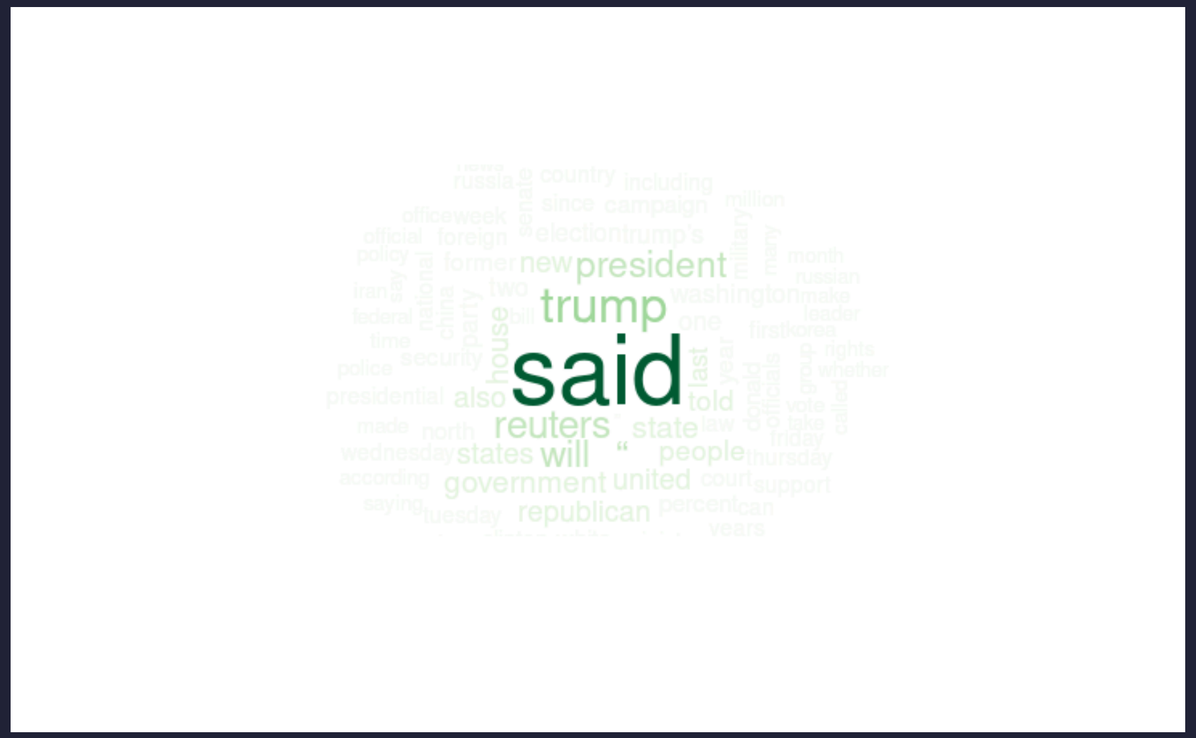
02 Frequent use of personal names and political terms shows media focus on individual-centric news.

03 High frequency of the word “said” implies attribution, which is common in real news.

04 Useful for developing a keyword-based feature extraction mechanism.

05 Encourages topic modeling or TF-IDF vectorization for deeper insights.

Wordcloud (True News)



01 Top terms include "said", "reuters", "president", "government", "state" – all formal and factual.

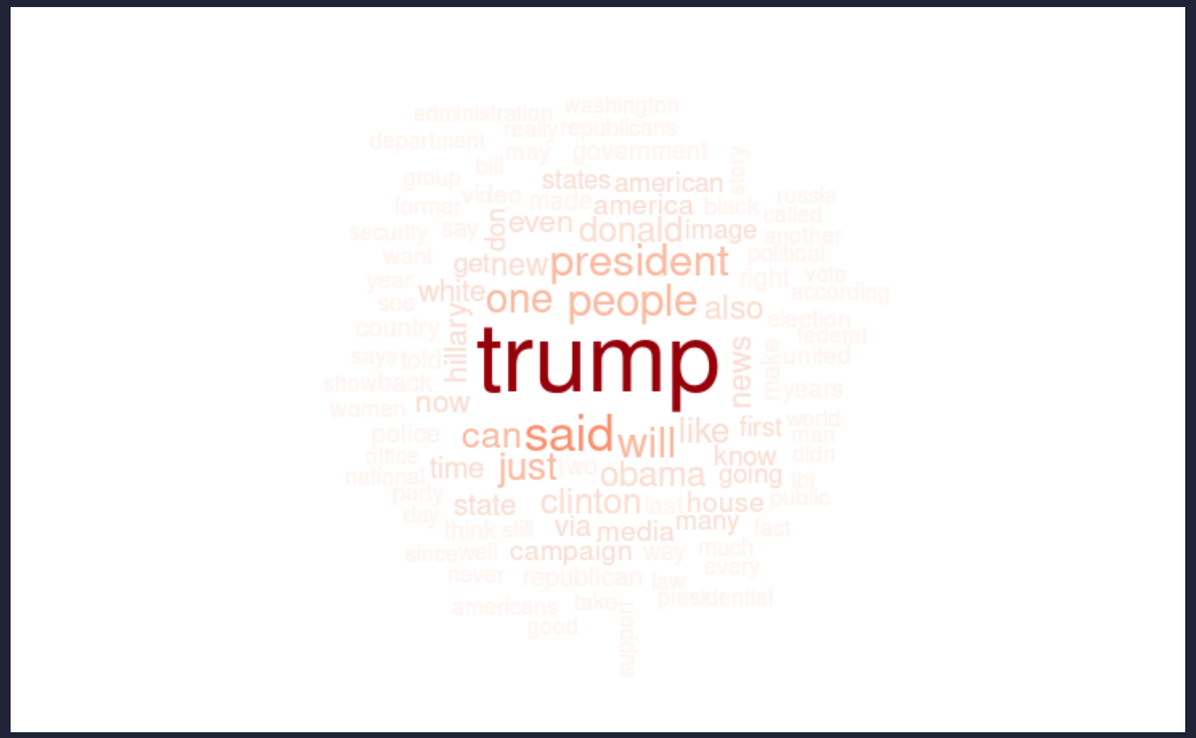
02 True news tends to cite sources and use specific terminology.

03 Terms like “states”, “republican”, “campaign”, “law” suggest structured reporting.

04 Emphasizes the presence of official and institutional language.

05 Highlights a clear linguistic contrast with fake news content.

Wordcloud (Fake News)



01 The term "Trump" is the most dominant, suggesting frequent mentions in fake news.

02 Other highlighted words like “president”, “people”, “can”, “said” indicate politically charged content.

03 Vocabulary used leans towards general, vague, and emotionally provocative language.

04 This reveals potential keyword patterns that could [illegible] classification of fake news.

05 The repetition of names such as "Clinton", "Obama", "Hillary" suggests fake news targets political figures disproportionately.

THANK YOU!