# Fake News Detection

## **UMASS Lowell**

[Names: AJ Pipattanakun, Sudhir Gunaseelan, Shuyu Lin]
[Student IDs: 01831790, 02003129, 01950812]

# Introduction & Motivation

**Problem Overview**:

- Fake news is a growing concern in the digital age, where misinformation spreads rapidly and influences public opinion and decision-making.
- Combating fake news requires effective tools that can distinguish between fake and real news with high accuracy.

**Significance**:

- Fake news detection is critical for maintaining trust in journalism and preserving the integrity of information in society.
- Beneficiaries include media organizations, fact-checking agencies, and the general public.

**Challenges**:

- Wide variations in the writing styles of fake and real news.
- Intentional obfuscation by fake news creators.
- Complexity of language processing and the need for scalable solutions.

**Solution Summary**:

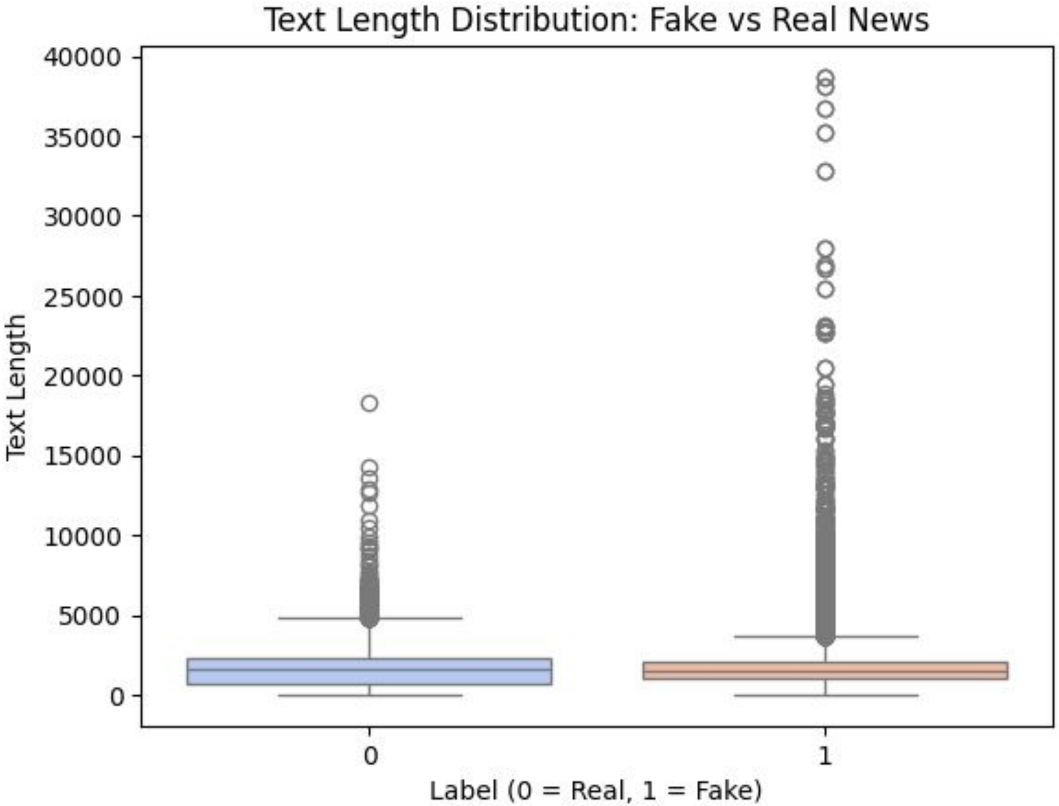
a. Our project employs machine learning models, leveraging features like term frequency-inverse document frequency (TF-IDF) to analyze news articles.
b. By training and evaluating multiple classifiers, we achieve reliable fake news detection.

# Dataset, Statistics & Evaluation Method

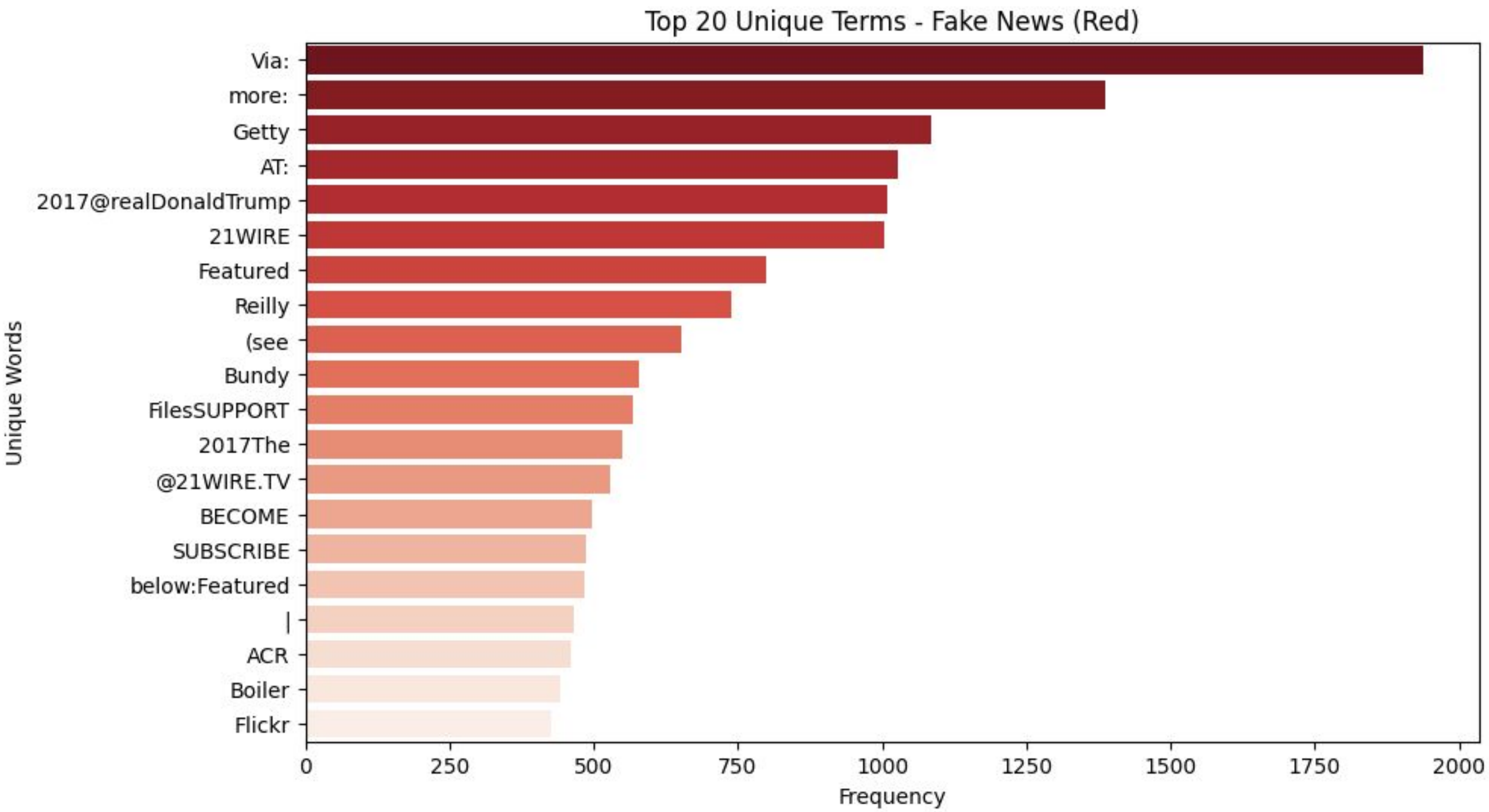
**Key Insights**:

- **Dataset Chosen:** *Fake and Real News Dataset* from kaggle
- **Data Analysis:**
  - Number of articles: 23502 fake, 21417 true.
  - Dataset Columns: Title, Text, Subject, Date.
- **Text Length Distribution**:
  - "Real news articles tend to have a higher average text length compared to fake news articles."
  - Boxplot visualization showing the differences.
- **Top 20 Unique Terms**:
  - "Fake news articles often use sensational and repetitive terms."
  - "Real news articles are characterized by neutral and formal language."
  - Bar charts for unique terms in fake and real news.
- **Common Words**:
  - "Word clouds reveal frequently used words in fake vs. real news articles."
  - Fake news: Sensational words like 'breaking', 'shocking'.
  - Real news: Formal words like 'government', 'official'.
- **Evaluation Metrics**:
  - Accuracy, Precision, Recall, F1 Score, and ROC-AUC were calculated to assess model performance.

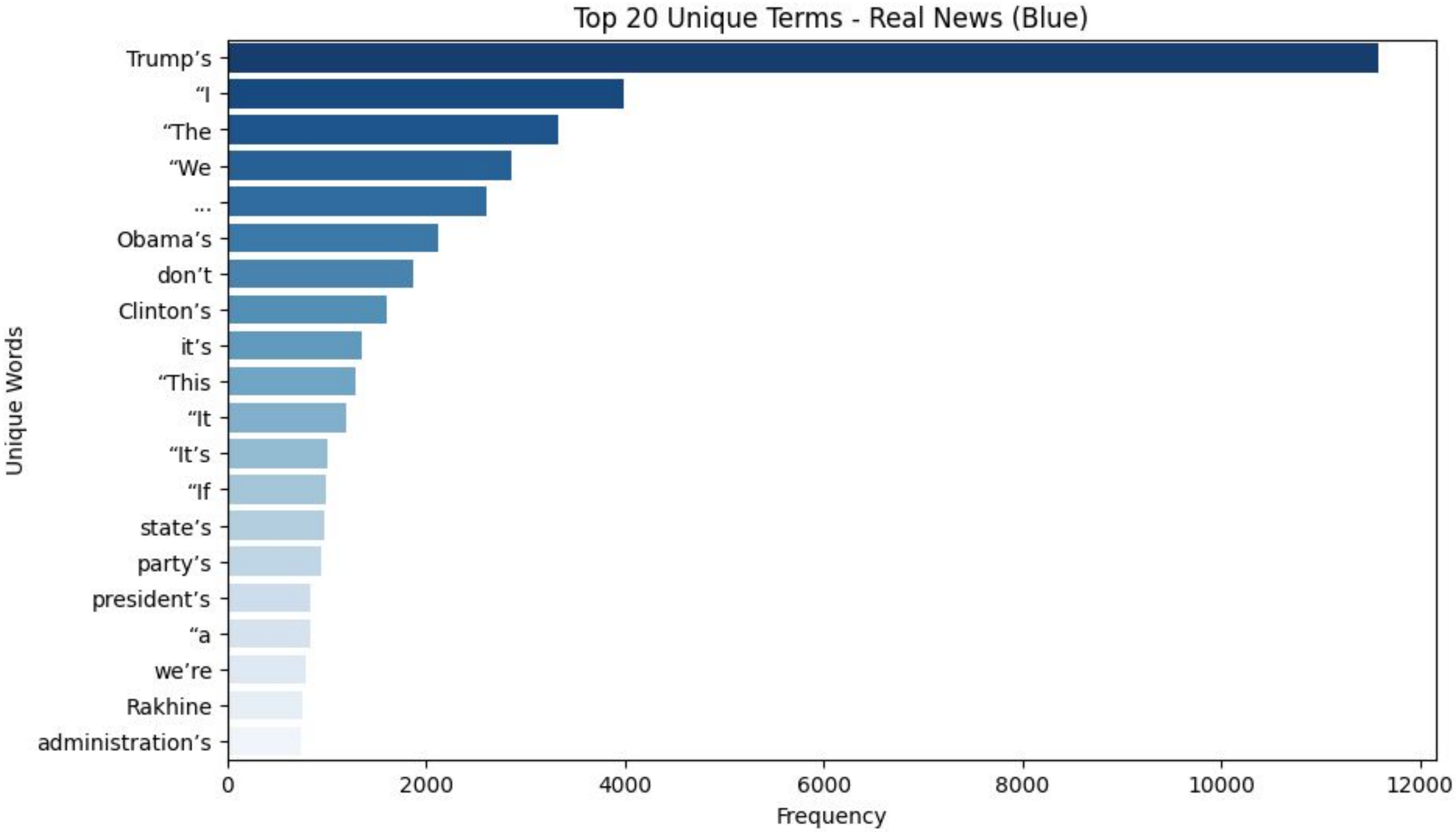
# Text Length Distribution



# Top 20 Unique Terms - Fake News



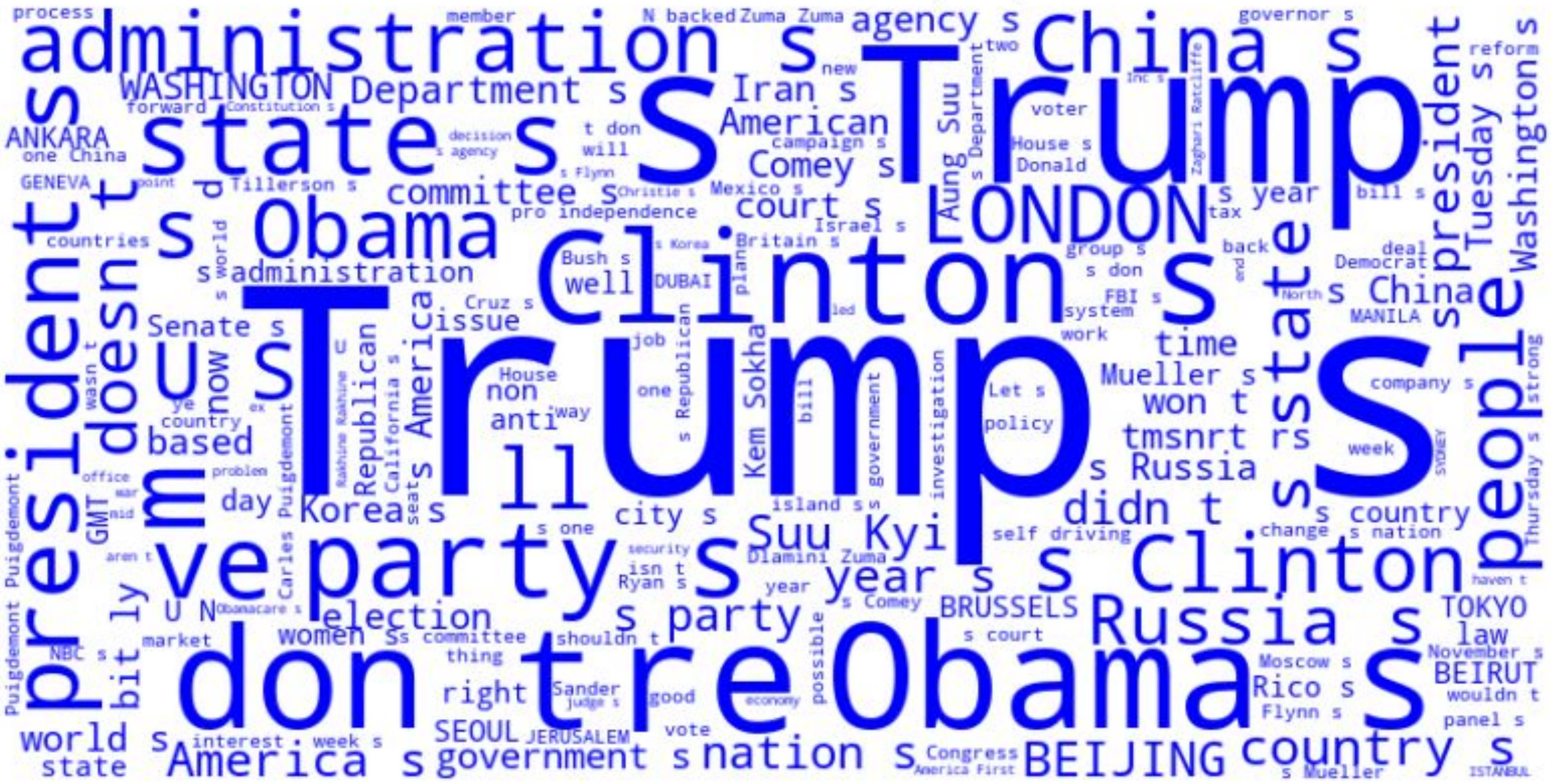
# Top 20 Unique Terms - Real News



# Word Cloud - Fake News

# Word Cloud - Real News

Word Cloud - Real News (Blue)



# Method - Model training

1. Logistic Regression
   a. Assuming linear relationship between features and the target.
   b. Faster and easier to tune but prone to overfitting.
   c. Efficient and fast but can overfit on small datasets.
2. Random Forests
   a. Capture the non-linear relationship between features and the target.
   b. Take longer to train but generalize better.
   c. More robust but requires more training time

**Training Pipeline**:

- Preprocessing: Text cleaning and tokenization.
- Feature Extraction: TF-IDF vectorization.
- Model Training: Logistic Regression and Random Forest.

# Results

Best Logistic Regression Model:

1. Accuracy: 0.9934298440979955
2. Precision: 0.9952728835410399
3. Recall: 0.9920753908759906
4. F1 Score: 0.9936715649469055
5. Confusion Matrix:
   [[4289   22],
   [37 4632]]

# Insights - Limitations & Potential Future Works

1. Overfitting: including more data from longer time period, the current model may be overfitting as there only 40000 training examples.
2. Could add pre-trained word embedding and use more complex model to help the model better understanding the context.
3. Could include more information from outside of journalism.
4. Simple model like logistic regression is better capturing the relationship among words and articles but cannot capture the complexity of a language.
5. Performance on synthetic data were not as great as the testing data.

# References


- **Dataset Reference:** *Fake and Real News Dataset* - Fake and Real News Dataset from kaggle