# Task 1: Fake News Detection

# Muhammad Sharique Baig

**1. Introduction**

The rapid spread of misinformation through online news portals and social media has made **fake news detection** a major research challenge. This project aims to build a **machine learning model** to classify news articles as either **fake** or **real** using the [Fake and Real News Dataset](https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset).

The dataset consists of news articles collected in the **United States from 2015 to 2018**, a politically charged period marked by the U.S. presidential election and its aftermath.

**2. Data Preprocessing**

* **Text Cleaning**: HTML tags, URLs, numbers, and punctuation were removed.
* **Stopwords**: Common words (e.g., the, is, am) were removed.
* **Lemmatization**: Words were reduced to their base form (e.g., running → run).
* **Combined Features**: The article title and text were merged for stronger features.

**3. Dataset Overview**

The dataset contained the following balance between fake and true articles:

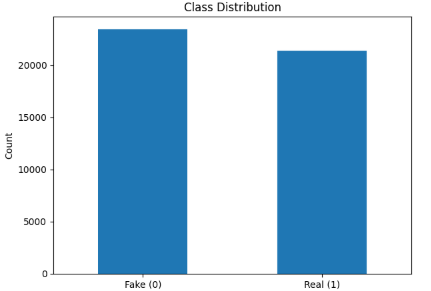


Figure Class Distribution

**4. Feature Engineering**

I used **TF-IDF Vectorization** with unigrams and bigrams (ngram\_range=(1,2)) to represent the articles

**5. Model Training & Evaluation**

I trained five machine learning models:

* Logistic Regression
* Multinomial Naïve Bayes
* Random Forest
* Gradient Boosting
* Linear SVM

Each model was evaluated using **accuracy, precision, recall, and F1-score**. The best model (based on F1-score) was selected as the final classifier.

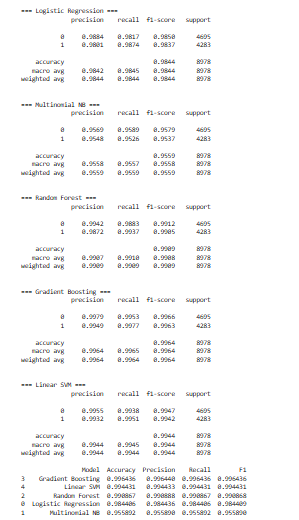


Figure Model Performance

Gradient Boosting was chosen and its performance is illustrated by the confusion matrix:

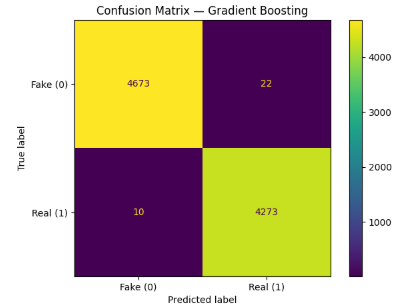
****

Figure Confusion Matrix

**6. Linguistic Analysis of Fake vs. Real News**

### 6.1 Top 20 Most Common Words

Fake and true news articles show clear differences in word usage.

* **Fake news** prominently featured words like trump, president, obama, clinton, hillary.
* **True news** relied on more institutional terms like said, u.s., government, republican, house.

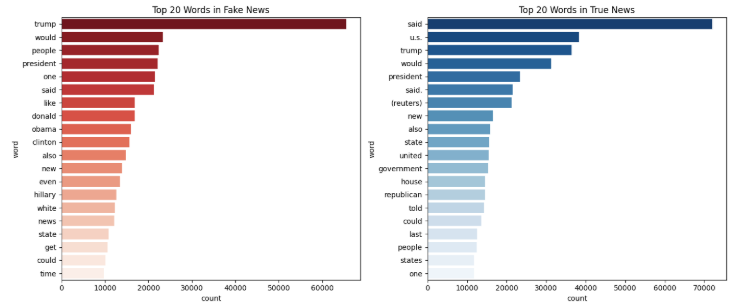
****

Figure Top 20 common words

### 6.2 Swear Words and Offensive Language

A striking difference emerged in the **use of profanity**:

* **Fake news articles contained 500+ swear words**.
* **True news articles contained only ~10 swear words** across all articles.

This highlights that **fake news tends to rely more on sensationalism, emotional appeal, and offensive language** compared to fact-based reporting.

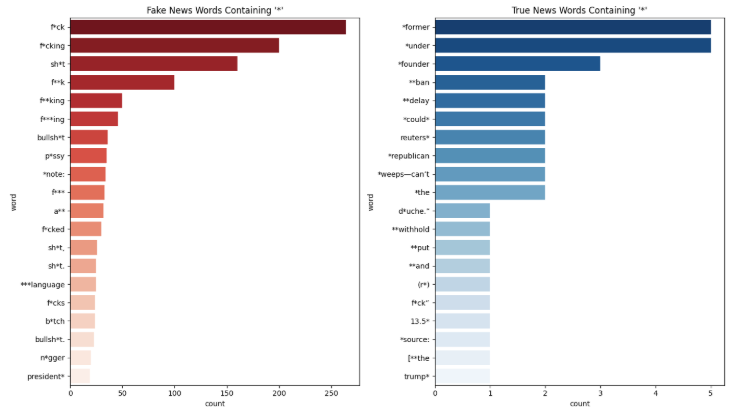
****

Figure Words containing \* (basic indicator of swear word)

### 6.3 Word Clouds

Word clouds illustrate the most frequent words in fake vs. true articles. Fake news clouds are dominated by **names of political figures**, while true news focuses on **reporting language** (said, state, u.s.).

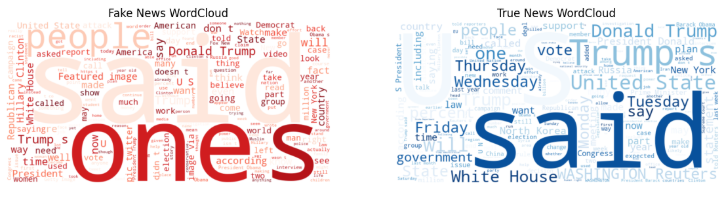
****

Figure Word Clouds

### 6.4 Article Length Distribution

Article length distributions also differ:

* **Fake news articles are often shorter**, suggesting quick, attention-grabbing content.
* **True news articles are generally longer**, consistent with detailed reporting.

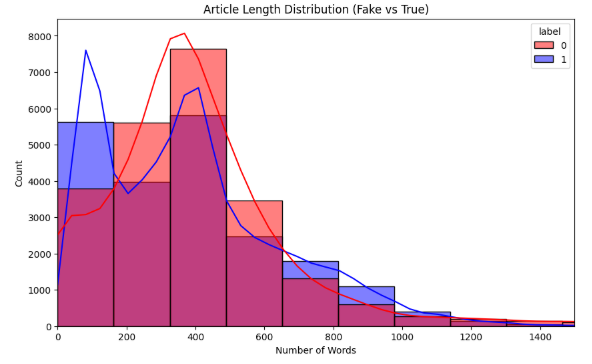
****

Figure Article Length

**7. Key Findings**

1. **Profanity as a Signal**: Fake news uses **50x more swear words** than true news.
2. **Different Lexical Styles**: Fake news emphasizes political figures and emotional hooks; true news emphasizes institutions and formal reporting.
3. **Article Length**: Fake news tends to be shorter.
4. **Model Accuracy**: The best-performing model (Gradient Boost) achieved strong accuracy and F1-scores, demonstrating the effectiveness of TF-IDF features.

**8. Streamlit App Integration (Work in progress)**

The trained TF-IDF vectorizer and best model were **saved using joblib** (tfidf\_vectorizer.joblib, best\_model.joblib). These can be loaded in a **Streamlit app** to:

* Input a news snippet.
* Predict whether it is **Fake (0)** or **Real (1)**.
* Display **highlighted keywords** (e.g., profanity or political names) that influenced the decision.

This makes the classifier **transparent and explainable**, showing users why an article was flagged as fake but it is yet to be implemented.

**9. Conclusion**

This project demonstrates that **machine learning can effectively detect fake news** by leveraging textual patterns. The results highlight not just model performance but also **linguistic differences** between fake and real news in the U.S. (2015–2018).

The findings reinforce the notion that **fake news relies more heavily on emotional triggers, profanity, and brevity**, whereas **real news relies on structured reporting and institutional credibility**.