Fake News Detection Using Machine Learning

Data Preprocessing, EDA, and Modeling

Jongwook Choe

San Francisco State University

December 8, 2025

Outline

1. Introduction
2. Data Preprocessing
3. Exploratory Data Analysis
4. Modeling Approach
5. Results
6. Conclusion

Introduction

Have you ever questioned whether the news you read is real? Since the rise of the internet, misinformation has become increasingly common and is often used to influence public opinion.

Example: Nayirah Testimony (1990)

- A 15-year-old Kuwaiti girl falsely testified that Iraqi soldiers removed babies from incubators.
- Her story increased public support for U.S. military intervention.
- Later proven false by ABC reporter John Martin (1991).

Motivation

This raises an important question: **What should we believe, and what should we doubt?**

In an age of rapid misinformation spread, distinguishing truth from falsehood is extremely difficult.

Project Goal:

- Build a machine learning model that predicts whether a news article is real or fake
- Use only **keyword frequency patterns** in the article's text
- Evaluate classical ML and deep learning approaches

Data Preprocessing

Dataset Used: BuzzFeed News

- 91 real news articles
- 91 fake news articles
- **182 total entries** after merging

Since the dataset is small, **all entries were retained** to maximize the amount of training data.

Preprocessing Steps:

- Extracted and merged labels from file names
- Removed irrelevant columns (URLs, images, metadata)
- Added binary content features:
  - `contain_movies`
  - `contain_images`

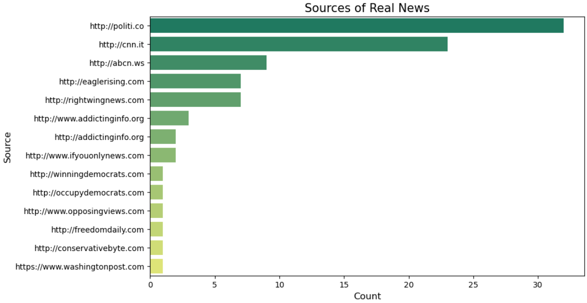
Text Cleaning Procedures

A custom text preprocessing pipeline was implemented to clean the article body text.

1. Convert all text to lowercase
2. Remove numbers
3. Remove punctuation
4. Remove special characters
5. Remove stopwords
6. Apply Porter stemming
7. Strip extra whitespace

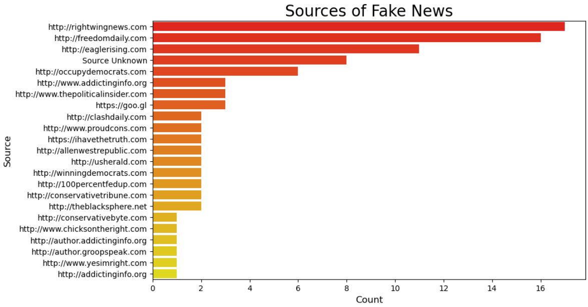
The final preprocessing function was used as the analyzer in the `CountVectorizer` to transform text into features.

EDA: Sources of Real News



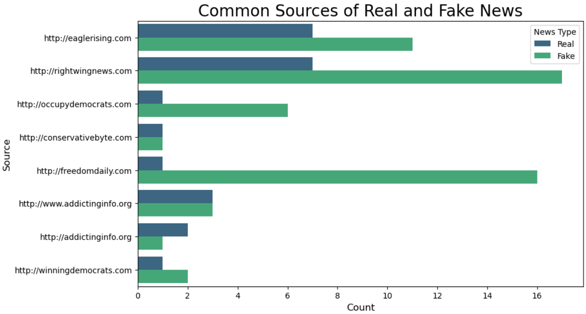
- A small number of sources dominate the real-news dataset:
  - `politi.co` and `cnn.it` appear most frequently.
  - A long tail of low-frequency sources contributes only 1–2 articles each.
- Distribution is **highly imbalanced** toward a few major outlets.

EDA: Sources of Fake News



- Fake news sources are more fragmented compared to real news:
  - `rightwingnews.com` and `freedomdaily.com` are the most common.
  - Many sources contribute only a single article.
  - Some entries show **“Source Unknown”**, indicating missing metadata.
- Fake news outlets show a **more decentralized distribution** than real news.

EDA: Common Sources in Real and Fake News



- Several domains appear in **both real and fake** subsets of the BuzzFeed dataset.
- `rightwingnews.com`, `eaglerising.com`, and `freedomdaily.com` show **significantly higher counts** in the fake news category.
- The overlap suggests that:
  - source alone is **not** a reliable indicator of truthfulness,
  - content-based features (text frequency, keywords) remain essential.

EDA: Distribution of Image and Video Links

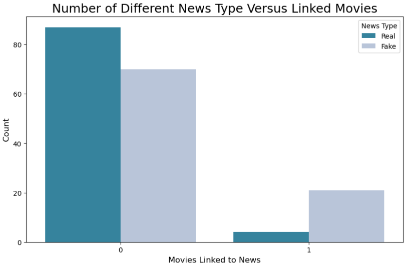
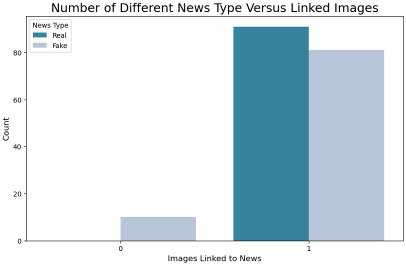


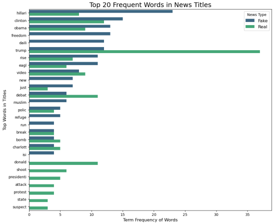
Image Link Distribution



Video Link Distribution

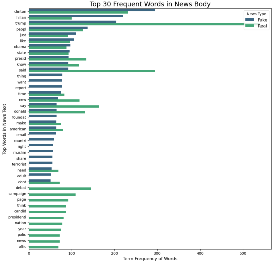
- Both image and video link counts show a **highly skewed distribution**.
- Most articles contain **zero** images or videos, while a few contain several.
- Although the raw counts are imbalanced, the presence/absence signal is still informative.
- Therefore, both were converted into **binary features**:
- These features were kept and used in the final model.

EDA: Top Words in News Titles



- **Fake News:** Features emotionally charged political names (`hillari`, `clinton`, `obama`) and ideological terms (`freedom`).
- **Real News:** Heavily dominated by `trump` (highest frequency overall) and related political figures (`donald`, `presidenti`).
- **Action:** This high separation enabled the creation of strong title-based binary features.

EDA: Top Words in Body Text



- **Real News:** `trump` is the single most frequent word by a large margin (over 500 mentions).
- **Fake News:** Characterized by high frequencies of `clinton`, `hillari`, and other associated political figures, alongside general reporting terms.
- **Note:** Frequencies in the body text are much higher than in titles, reinforcing the importance of text-based indicators.

Modeling Approach

Feature Engineering:

- Extracted top frequent **fake/real words** from:
  - news titles
  - news body text
- Created binary indicator columns:
  - `fake_title_word`, `real_title_word`
  - `fake_body_word`, `real_body_word`
- Encoded target using `LabelEncoder`

Models Evaluated:

- Logistic Regression
- Random Forest (100 trees)
- Gradient Boosting (100 estimators)
- Bagging Classifier (100 decision trees)
- K-Nearest Neighbors ($k = 5$)

Feature Selection (SFS)

Method: Sequential Forward Selection (SFS)

- Logistic Regression used as base estimator
- Selected the top 10 most predictive engineered features

Selected Features Included:

- `contain_movies`, `contain_images`
- Fake-title indicators:
  - `fake_title_hillari`
  - `fake_title_clinton`
  - `fake_title_obama`
  - `fake_title_freedom`
  - `fake_title_daili`
- Real-title indicators:
  - `real_title_trump`
  - `real_title_clinton`
  - `real_title_donald`

Result: Feature-selected model and full model gave **the same accuracy** (no improvement due to small dataset size).

Model Accuracy Comparison

Accuracy of All Models on Test Set

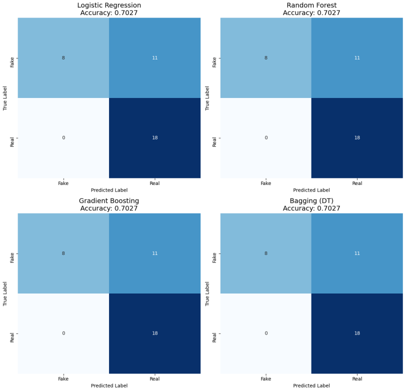
Model	Accuracy
Logistic Regression	0.702703
Random Forest	0.702703
Gradient Boosting	0.702703
Bagging (DT)	0.702703
KNN (k = 5)	0.513514

Key Takeaways:

- All four models achieved the same accuracy of **0.7027**.
- KNN performed significantly worse at **0.5135**.
- The uniform performance suggests:
  - limited dataset size (182 samples), and
  - engineered features dominate model performance.

Confusion Matrix Comparison

Confusion Matrices Across Models



All four top-performing models (LR, RF, GB, Bagging) produced **identical** confusion matrices.

Confusion Matrix Interpretation

Key Insights:

- **Perfect Real News Detection:**
  - 18/18 Real articles correctly classified
  - Recall (Real) = **100%**
- **Errors Come From Fake News:**
  - 11 Fake articles misclassified as Real (False Positives)
  - Only 8 Fake articles correctly detected
  - Recall (Fake) = **42%**
- **Main Takeaway:** Models easily recognize Real News but struggle to correctly identify Fake News.

Optimizing KNN Hyperparameters

Grid Search Setup:

- Tested $k = 1$ to $k = 20$
- Compared `uniform` vs. `distance` weights
- Stratified 5-fold cross-validation

Best Parameters Identified:

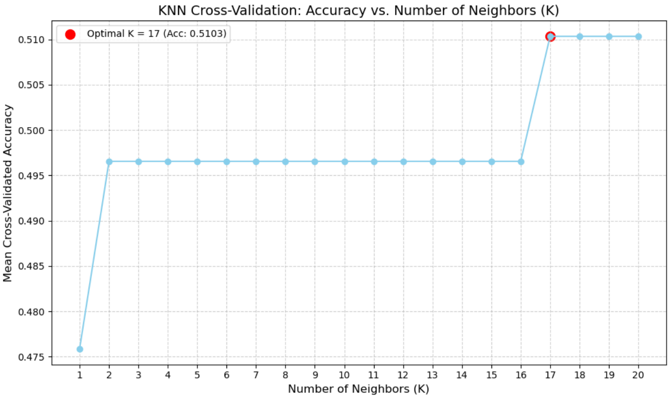
- Optimal k: **17**
- Best weighting: **uniform**
- Cross-validated accuracy: **0.5103**

Final Test Accuracy (Optimized KNN):

0.513

Due to sparse binary features, KNN does not perform well on this dataset.

KNN Performance Across Different Values of *k*



Key Observation:

- The accuracy curve remains low for all tested k values.
- Even at optimal $k = 17$, performance stays around 0.51.
- Confirms that KNN is not suitable for this feature space.

Conclusion

- Simple keyword-based engineered features can provide a reasonable baseline for fake news detection.
- Multiple ML models (LR, RF, GB, Bagging) achieved the **same accuracy**, showing that performance is limited by dataset size and feature richness rather than model choice.
- Models were highly effective at identifying Real News, but struggled to correctly classify Fake News due to overlapping vocabulary and limited examples.
- With only 182 articles, the dataset is too small to capture meaningful linguistic patterns—future work requires **larger, more diverse datasets**.
- Incorporating richer textual features (TF-IDF, embeddings, or deep learning models) could significantly improve detection ability.

References

- Opinion — *Remember Nayirah, Witness for Kuwait?* The New York Times, 1992. Available at: www.nytimes.com/1992/01/06/opinion/remember-nayirah-witness-for-kuwait.html Accessed 6 May 2025.
- Shu, Kai, et al. "*FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media*." arXiv, 2019. Available at: arxiv.org/abs/1809.01286.
- Mahudeswaran, Deepak. "*FakeNewsNet*." Kaggle, 2018. Available at: www.kaggle.com/datasets/mdepak/fakenewsnet/data.