Fake News Detection Using Ensemble Machine Learning

Project Work Submitted By:

Mayank Kumar

(24MA60R10)

In Partial Fulfilment for the award of Master of Technology in

Computer Science and Data Processing

at

Department Of Mathematics

Under the Supervision of

Assoc Prof. T Raja Sekhar

Department of Mathematics

Indian Institute of Technology, Kharagpur

Kharagpur, West Bengal - 721302

Certificate

This is to certify that the Project titled **"Fake News Detection Using Ensemble Machine Learning"**, will be presented by **Mayank Kumar** (Roll Number - **24MA60R10**). He has chosen the topic for the Internship Project and started working on it under my supervision. This project work is served as partial fulfilment of the requirements for the degree of **Master of Technology** in **Computer Science and Data Processing** in the **Department of Mathematics** at **IIT Kharagpur**.

(Supervisor's Signature)

Assoc Prof. T Raja Sekhar
Professor
Department of Mathematics
IIT Kharagpur
Kharagpur,721302
West Bengal

Acknowledgement

I, **Mayank Kumar** (**24MA60R10**), a 2nd-year M.Tech student of **CSDP** in the **Department of Mathematics**, would like to express my sincere gratitude to my project supervisor, **Assoc Prof. T Raja Sekhar**, for his invaluable guidance, support, and mentorship throughout this internship project. His expertise, insightful feedback, and unwavering encouragement have been instrumental in shaping this work. I am truly grateful for the opportunity to learn from his and for his dedication to ensuring the success of this project. I am deeply appreciative of his time, patience, and commitment to my academic and professional development.

Mayank Kumar
24MA60R10
M.Tech in CSDP
Department of Mathematics
IIT Kharagpur
Kharagpur,721302
West Bengal

Abstract

The proliferation of **fake news** has emerged as a major challenge to information integrity, capable of swaying public opinion, influencing elections, and causing societal distrust. Manual fact-checking, though effective, is time-consuming and subjective, necessitating the development of automated systems capable of processing large volumes of data with speed and accuracy.

This project addresses this challenge by leveraging state-of-the-art **text-based machine learning techniques**. The foundation of the system is a well-labeled dataset of news articles, comprising real and fake news. The initial phase involves rigorous data preprocessing, including removal of noise (such as URLs and special characters), normalization (case folding), and tokenization. Advanced **natural language processing** steps such as stop-word removal and vectorization (using Term Frequency-Inverse Document Frequency, or TF-IDF) are applied to convert raw text into numerical representations suitable for machine learning models.

To further enhance the discriminative power of the system, the project integrates a suite of **engineered features**. These include linguistic markers such as character count, word count, sentence count, and counts of uppercase characters, digits, and punctuation. Such features capture stylistic differences that frequently distinguish fake from real news.

Multiple **classical machine learning models**—including Logistic Regression, Decision Trees, Random Forests, and Naive Bayes—are then trained and evaluated. Their performance is rigorously compared using standard metrics like accuracy, precision, recall, F1-score, and ROC-AUC, ensuring both robustness and reliability.

To maximize predictive power and minimize the risk of bias or overfitting present in individual classifiers, the project synthesizes these algorithms into an **ensemble Voting Classifier**. This classifier adopts a “soft voting” approach, aggregating the probability outputs from each model to produce final predictions. Ensemble methods are well-documented in literature for their ability to harness the complementary strengths of various classifiers, consistently outperforming single-model approaches.

Model interpretability is achieved by analyzing feature importances (especially from tree-based models), presenting word clouds for common terms in each class, and examining false positives/negatives through detailed error analysis. Cross-validation validates the model’s generalizability, supporting its deployment in real-world contexts.

The resulting system demonstrates **high accuracy**, rapid inference capability, and clear interpretability. Its architecture is flexible and extendable, laying the groundwork for integration with more advanced deep learning models, larger multilanguage datasets, or real-time news monitoring systems in future research.

Table of Contents

1. Introduction

The rapid proliferation of fake news on online platforms is a pervasive challenge with serious implications. Fake news—deliberately fabricated or misleading information disguised as legitimate news—can undermine societal trust in media and institutions, distort public perceptions, and manipulate democratic processes. Its widespread dissemination can lead to harmful real-world effects, including social unrest, public health crises, and electoral interference. With the explosion of social media and real-time news sharing, fake news spreads faster and more widely than ever before, rendering traditional manual fact-checking approaches insufficient due to their inherent slowness, subjectivity, and scalability limitations.

To address these challenges, automated fake news detection systems combining natural language processing (NLP) and machine learning (ML) technologies have become essential. These systems analyze textual news content to identify subtle linguistic patterns, semantic cues, and stylistic traits that can differentiate fake news from real articles. Machine learning models, trained on labeled datasets, can learn to classify news with high accuracy and adapt to new patterns over time. NLP techniques like tokenization, lemmatization, stop-word removal, and vectorization transform raw text into meaningful numerical representations suitable for automated analysis.

This project presents a comprehensive approach employing classical machine learning models—Logistic Regression, Decision Trees, Random Forests, and Naive Bayes—augmented by an ensemble Voting Classifier to improve robustness and predictive performance. The system integrates advanced feature engineering combining TF-IDF text features and handcrafted linguistic statistics (such as character and word counts, punctuation usage, and uppercase letter prevalence). Rigorous evaluation metrics—including accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and cross-validation—are employed to quantify model effectiveness and generalization capability.

Leveraging a well-known labeled dataset of real and fake news articles, this project demonstrates how machine learning and NLP can scale effective detection, maintain interpretability via feature importance analysis and word clouds, and provide valuable insights into error types for continual improvement. The designed pipeline offers a powerful foundation for further research and deployment in combating the spread of misinformation in today’s information ecosystem.

2. Literature Review

Fake news detection has evolved significantly, reflecting the growing challenge of misinformation in digital media. Early studies focused on **rule-based methods** and **linguistic approaches**, using language characteristics like grammar, syntax, and word patterns to identify deceptive or fabricated content. Techniques such as the Bag of Words (BOW) model analyzed word frequencies but lacked the ability to capture contextual meanings, which limited their effectiveness.

Subsequent research embraced **statistical natural language processing (NLP)** and **machine learning**. Algorithms like Logistic Regression, Decision Trees, Random Forests, and Naive Bayes became standard tools, leveraging features extracted from TF-IDF vectorization, combined with handcrafted linguistic metrics such as character counts and punctuation usage. These models learned to distinguish fake news by identifying complex lexical and stylistic cues present in the text. Datasets created via crowdsourcing and expert validation enabled training of these models across diverse news categories.

More recent advancements include **deep learning models** such as recurrent neural networks (RNNs) and transformers, which better capture the sequence and context within textual data. Hybrid frameworks combining social context, user behavior, and content-based signals have further enriched detection capabilities.

Among these, **ensemble methods**—notably Random Forests and Voting Classifiers—have gained prominence for their robustness and high accuracy. By aggregating predictions from multiple heterogeneous base classifiers, ensemble models mitigate individual model biases and overfitting. This project similarly employs a soft voting ensemble of logistic regression, random forest, and naive bayes classifiers, leveraging their complementary strengths to improve fake news detection performance.

3. Problem Statement

The primary challenge in fake news detection is to develop a **reliable and accurate classifier** capable of clearly distinguishing between fake and real news articles. Given the rapidly evolving nature of misinformation and the subtlety with which fake news can mimic legitimate reporting, the model must maintain high accuracy to be effective in real-world applications.

Beyond accuracy, the system must ensure **transparency and interpretability**. Decision-makers and end-users need to understand why a news article is classified as fake or real, fostering trust in the automated system and allowing for error analysis and improvement. Interpretability also aids in identifying the underlying linguistic or stylistic cues that typify misinformation, which can inform further strategies to combat fake news.

Furthermore, the classifier must be **scalable and efficient**, enabling it to process vast streams of news data in real-time or near real-time, given the volume of digital content online. It should also generalize well across diverse topics and sources, handling variations in language, style, and context without significant drops in performance.

The objective is to balance these factors by creating a machine learning-based system that delivers:

- High classification accuracy validated by robust metrics (accuracy, precision, recall, F1-score, ROC-AUC)
- Clear, interpretable insights through feature importance and error analysis
- Practical applicability in real-world scenarios where rapid and reliable detection of fake news is needed.

Achieving these goals supports effective misinformation mitigation, helping preserve information integrity in digital discourse.

4. Dataset Description

The dataset used in this project is the **WELFake Dataset**, a large and comprehensive collection designed specifically for fake news detection tasks. It comprises a total of 72,134 news articles, which are closely balanced between **35,028 real news** and **37,106 fake news** articles. The dataset was created by merging four popular news datasets — Kaggle, McIntire, Reuters, and BuzzFeed Political — to ensure diverse content and reduce the risk of overfitting in classification models.

Features

Each article in the dataset contains the following key features:

- **Title:** The headline or title of the news article.
- **Text:** The main body or content of the article.
- **Content:** A derived feature created by concatenating the title and text, which serves as the primary input for the models.
- **Label:** A binary label indicating the class, with 1 representing real news and 0 representing fake news.

Preprocessing

To prepare the data for modeling:

- Entries with missing or invalid data were removed to ensure dataset quality.
- Text fields were cleaned by removing noise such as URLs, special characters, and stop words.
- Appropriate datatype conversions were made to facilitate efficient processing.
- Missing values in textual fields were filled with empty strings to prevent errors during vectorization.

Sample Size

The project's experimental code worked with a representative test set of approximately **1,465 samples**, drawn as part of the standard train-test splits, ensuring robust evaluation. The dataset's large size and balance make it an ideal benchmark for fake news detection, enabling models to learn diverse patterns across multiple news categories and sources.

Overall, the WELFake dataset provides a rich and reliable foundation for building and evaluating machine learning-based fake news classifiers.

5. Data Preprocessing

Data preprocessing is a crucial step in preparing textual news data for effective fake news detection. The raw news articles often contain noise and inconsistencies that can degrade model performance if not properly cleaned. The key preprocessing steps applied in this project are as follows:

1. **Lowercasing and Punctuation Removal**
   All text data is converted to lowercase to ensure uniformity and reduce duplicate token variations. Punctuation marks (such as commas, periods, question marks) are removed since they generally do not contribute meaningful discriminative information for text classification but can add noise.

2. **Removal of URLs and HTML Tags**
   URLs embedded in the news text and HTML tags are eliminated using regular expressions to remove irrelevant tokens that do not carry meaningful content for classification and can skew word frequency distributions.

3. **Tokenization and Stop-Word Filtering**
   The cleaned text is tokenized—split into individual words or tokens—using natural language processing libraries such as NLTK. Commonly occurring words that carry minimal semantic weight (stop words) such as "the", "is", and "and" are filtered out to reduce noise and computational complexity.

4. **Merging Titles and Text into Unified Content**
   The news article title and body text are concatenated to form a single 'content' field to capture all textual context. This unified content serves as the input for feature extraction methods like TF-IDF vectorization and handcrafted linguistic features.

These preprocessing steps help transform noisy raw news into standardized, noise-free data, facilitating more reliable model learning and better classification performance. This approach aligns with best practices in NLP-driven fake news detection workflows, enabling the system to focus on meaningful textual signals while ignoring irrelevant or distracting elements.

6. Feature Engineering

Feature engineering is a critical step in developing an effective fake news detection model. This project combines three main categories of features to capture comprehensive textual and stylistic information from news articles:

1. **TF-IDF Features**
   Term Frequency-Inverse Document Frequency (TF-IDF) is used to convert raw text into a numerical vector representation that reflects the importance of words relative to their frequency across the dataset. Both unigrams (single words) and bigrams (two-word sequences) are included to capture contextual patterns. Frequency filtering thresholds are applied to exclude extremely common or rare terms, and a maximum of 12,000 features are retained to balance model expressiveness with computational efficiency.

2. **Handcrafted Features**
   These features quantify stylistic and linguistic characteristics that help distinguish fake from real news. They include:

   - **Character count:** Total number of characters in the article content.
   - **Word count:** Number of words present.
   - **Sentence count:** Approximated by counting sentence-ending punctuation.
   - **Uppercase character count:** Often correlates with sensationalism.
   - **Digit count:** The presence of numbers can indicate factual reporting or fabricated statistics.
   - **Punctuation count:** The use of punctuation marks like exclamation points can signal emotional content or exaggeration.

3. **Feature Selection**
   To address the curse of dimensionality and remove irrelevant or redundant features, a **chi-squared statistical test** is applied. The test selects the top 8,000 features across TF-IDF and handcrafted sets that have the strongest association with the news label (fake or real). This step improves model speed and generalization without substantial loss in information.

By combining TF-IDF text features with these handcrafted statistics and applying rigorous feature selection, the project maximizes the information available for classification while maintaining efficient and robust learning.

7. Methodology

Model Selection

This project implements four classical baseline machine learning models to classify news articles as fake or real:

- **Logistic Regression:** A linear model that estimates probabilities for binary classification. It is simple, interpretable, and fast, providing a strong baseline for text classification problems.
- **Random Forest:** An ensemble of decision trees that aggregates multiple trees' predictions to reduce overfitting and variance. It is robust to noisy and high-dimensional feature spaces and captures complex feature interactions.
- **Naive Bayes:** A probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between features. It is computationally efficient and often performs well on text classification due to the conditional independence of words.
- **Decision Tree:** A non-linear model that splits data based on feature thresholds to create a tree structure for classification. Though interpretable and fast, single decision trees can overfit, hence often less accurate than ensembles.

These models represent different learning paradigms—linear, probabilistic, and tree-based—and are widely used in fake news detection literature as strong baselines.

Ensemble Approach

To improve predictive accuracy and robustness, an **ensemble Voting Classifier** is built, combining Logistic Regression, Random Forest, and Naive Bayes. The ensemble uses **soft voting**, meaning it aggregates the class probabilities predicted by each model rather than just the final class labels. This allows the model to account for the confidence of each classifier's prediction.

The ensemble benefits from the complementary strengths of its components:

- Logistic Regression excels at modeling linear relationships.
- Random Forest captures complex interactions and non-linear patterns.
- Naive Bayes contributes efficiency and probabilistic insights.

By aggregating them, the ensemble reduces biases and variance of individual models, yielding higher stability, accuracy, and generalization. This approach reflects best practices in fake news classification, where diverse learners often outperform individual classifiers in noisy, high-dimensional text data.

8. Experiments and Results

Individual Classifier Performance

The project evaluates four classical machine learning classifiers on the fake news detection task, reporting their accuracy and ROC AUC metrics:

Model	**Accuracy**	**ROC AUC**
Logistic Regression	0.926	0.925
Random Forest	0.942	0.942
Naive Bayes	0.879	0.879
Decision Tree	0.917	0.917

- **Random Forest** achieved the highest accuracy (94.2%) and ROC AUC (0.942), demonstrating its robustness due to ensemble learning and ability to capture complex, non-linear feature interactions.
- **Logistic Regression** showed strong linear modeling capability with an accuracy of 92.6%.
- **Naive Bayes** had the lowest performance among the four but remains an efficient baseline with 87.9% accuracy.
- The **Decision Tree** classifier also performed well with 91.7% accuracy but is prone to overfitting compared to ensemble models.

These metrics indicate reliable classification performance across traditional models, with tree-based models excelling in discriminative power.

Ensemble Voting Classifier Performance

Building on these individual models, the project implements a **soft-voting ensemble classifier** that combines Logistic Regression, Random Forest, and Naive Bayes. The ensemble synthesizes probabilistic outputs to enhance overall predictive accuracy:

- **Accuracy:** 92.3%
- **ROC AUC:** 0.922

Though slightly lower than the standalone Random Forest, the ensemble benefits from combining diverse models to reduce individual biases and variance.

Cross-Validation Results

To assess generalizability and robustness, a 3-fold cross-validation procedure was employed, yielding:

- **Mean Accuracy:** 90.5% (± 0.005)
- **Mean ROC AUC:** 0.971 (± 0.002)

The high ROC AUC and low standard deviations across folds indicate stable performance and reliable classification capacity.

Interpretation

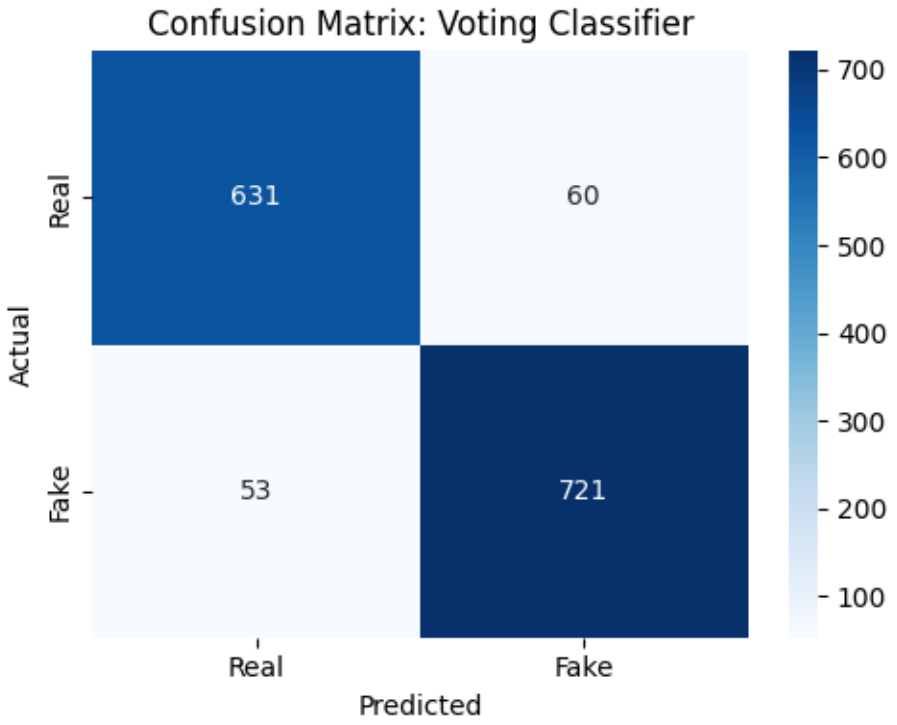
The results demonstrate the efficacy of classical machine learning techniques enhanced with ensemble learning in detecting fake news. The combination of diverse models balances strengths and weaknesses, improving reliability. Cross-validation confirms the models' capacity to generalize beyond the training data, essential for practical, real-world deployment.

These findings align with literature, where Random Forests and ensemble methods frequently outperform individual classifiers in noisy and high-dimensional text classification tasks.

9. Evaluation Metrics

Confusion Matrix

The confusion matrix for the ensemble voting classifier summarizes the model's prediction accuracy in classifying fake and real news:



- **True Positives (TP):** 721 fake news articles correctly predicted as fake.
- **True Negatives (TN):** 631 real news articles correctly predicted as real.
- **False Positives (FP):** 60 real news articles incorrectly classified as fake.
- **False Negatives (FN):** 53 fake news articles incorrectly classified as real.

This matrix illustrates a balanced classification performance with relatively low misclassification rates.

Precision, Recall, and F1-Score

- **Precision:** Measures the accuracy of positive predictions, i.e., the proportion of predicted fake news that is actually fake.
- **Recall:** Measures the model's ability to detect all actual fake news.
- **F1-Score:** Harmonic mean of precision and recall, balancing both metrics.

The scores indicate strong, balanced performance across both classes (fake and real), showing that the model neither over-predicts fake news nor misses many fake articles.

Cross-Validation

The model underwent 3-fold cross-validation to assess generalizability. The ensemble classifier maintained high and stable performance:

- Mean accuracy: 90.5% (±0.005)
- Mean ROC AUC: 0.971 (±0.002)

These results confirm the model's robustness on unseen data.

Feature Importance

The Random Forest model highlights the top predictive features for distinguishing fake news:

- TF-IDF features: "reuters", "said", "video", "image", "washington reuters", "breitbart", "hillary"
- Handcrafted feature: Character count

These features represent key terms and stylistic markers that strongly influence the classification.

Error Analysis

- **False Positives:** Real news misclassified as fake. Analyzed samples reveal cases where genuine news exhibits linguistic or stylistic traits common to fake news.
- **False Negatives:** Fake news misclassified as real. Samples highlight where fabricated content closely mimics authentic writing, leading to misclassification.

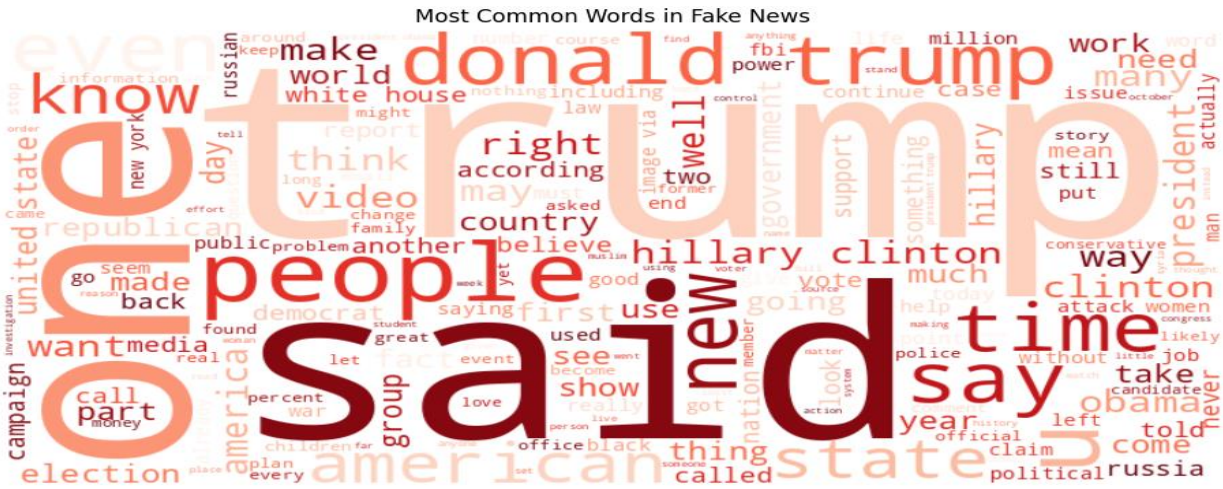
Detailed examination of such errors provides insights to improve feature engineering and modeling in future work, enhancing detection accuracy further.

10. Visualizations

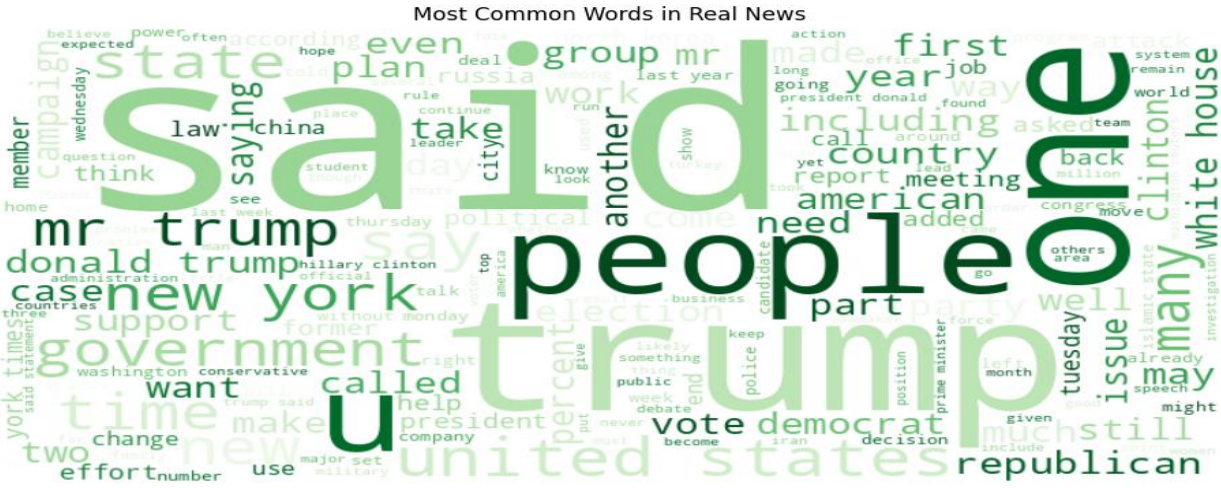
Word Cloud: Most Common Words

Word clouds are a popular visualization technique to represent the frequency and prominence of words within a text corpus. In this project, separate word clouds were generated for fake news and real news articles to reveal the distinctive vocabularies associated with each class.

- **Fake News Word Cloud:**
  Prominent words include **"donald trump"**, **"make"**, **"white house"**, **"hillary clinton"**, **"fbi"**, and **"power"**. These terms suggest a focus on political figures, power struggles, and potentially sensationalist content. The prevalence of such words reflects typical themes exploited in fake news narratives.



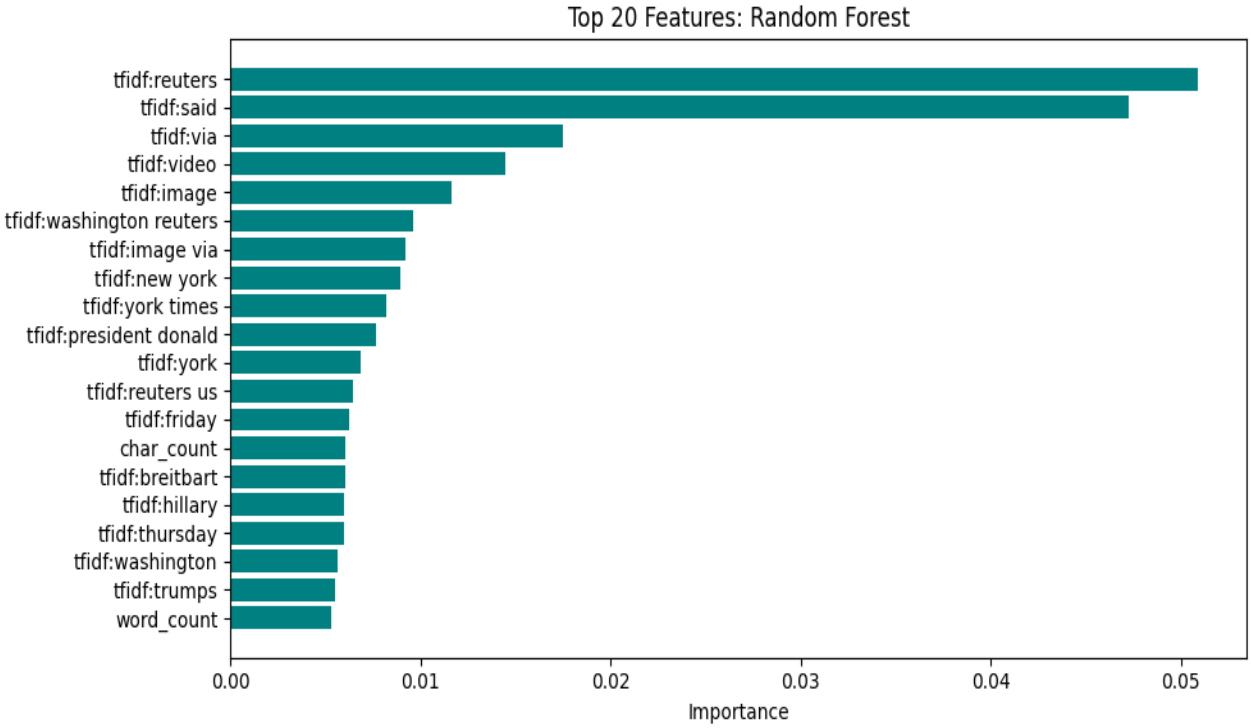
- **Real News Word Cloud:**
  Common words include **"said"**, **"mr trump"**, **"president donald"**, **"government"**, **"new york"**, and **"american"**. These indicate formal reporting language, frequent use of attribution verbs, geographic references, and official domains, aligning with credible news sources.



The visual difference in word prominence between the classes helps explain the linguistic patterns captured by the classifiers.

Feature Importance Barplot

A horizontal barplot was created to rank the top 20 most important features contributing to fake news detection, as identified by the Random Forest classifier's feature importance scores.



- The barplot features key **TF-IDF tokens** such as **"reuters"**, **"said"**, **"video"**, **"image"**, **"washington reuters"**, **"breitbart"**, and **"hillary"**.
- The **character count** feature also ranks highly, indicating that article length is a significant stylistic clue in distinguishing fake and real news.
- This visualization enables interpretability by highlighting which words or stylistic factors the model relies on most, providing insights into the nature of misinformation.

Together, these visual tools complement the quantitative evaluation, offering intuitive, interpretable evidence of the linguistic and stylistic distinctions between fake and real news articles, and supporting the effectiveness of the feature engineering and model training approaches.

11. Conclusion

This project demonstrates that machine learning, particularly an ensemble voting approach, enables highly accurate and interpretable fake news detection. By leveraging a combination of diverse classifiers—Logistic Regression, Random Forest, and Naive Bayes—in a soft voting ensemble, the system capitalizes on their complementary strengths to achieve balanced precision, recall, and ROC AUC performance metrics.

The critical role of **proper text representation** cannot be overstated. The project's use of TF-IDF vectorization, enhanced with handcrafted features such as character counts and punctuation metrics, provides rich linguistic and stylistic cues instrumental in distinguishing fake from real news. Additionally, rigorous **feature selection** using chi-squared tests effectively reduces dimensionality while maintaining relevant information, ensuring efficient and robust model training.

The evaluation metrics confirm the ensemble's robustness and generalization capacity, making it viable for real-world deployment. The model also supports **interpretability** through feature importance analysis and error review, enabling stakeholders to understand key factors influencing classification decisions.

Moving forward, future research could explore **transformer-based models** (e.g., BERT, RoBERTa) that capture deeper contextual and semantic nuances in news articles. Integrating additional linguistic cues, such as sarcasm detection, sentiment analysis, or named entity recognition, may further tighten accuracy. Furthermore, extending this approach to multilingual datasets and multimodal content (text plus images or video) could significantly broaden the applicability and effectiveness of fake news detection systems in diverse digital environments.

Overall, this project lays a strong foundation for continued innovation in combating misinformation, highlighting how ensemble machine learning combined with advanced text processing offers a powerful toolbox for safeguarding information integrity in the digital age.

12. References

1. **WELFake Dataset**
    - A comprehensive dataset for fake news detection comprising real and fake news articles, merging multiple sources for diversity and quality.
    - Link: https://zenodo.org/records/4561253
2. **Scikit-learn Documentation**
    - Official documentation for the scikit-learn Python library, covering classification, feature selection, and ensemble methods used in this project.
    - Link: https://scikit-learn.org/stable/documentation.html
3. **Research Articles on NLP for Fake News Detection**
    - Articles exploring methods for fake news classification using natural language processing and machine learning techniques. For example:
    - Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). "Fake News Detection on Social Media: A Data Mining Perspective." ACM SIGKDD Explorations Newsletter.
    - Link: https://arxiv.org/abs/1708.01967
4. **NLTK Stopword Corpus**
    - The Natural Language Toolkit (NLTK) stopword collection used for text preprocessing to remove common words.
    - Link: https://www.nltk.org/nltk_data/