FINANCIAL NEWS AUTHENTICITY PREDICTOR

Submitted by: Akshara V, Hariesh R, Lithikha B

SUMMARY

The **Fake Finance News Detection** platform tackles the growing challenge of misinformation in financial markets with a powerful machine learning solution that classifies news articles as real or fake. It leverages **FinBERT**, a model specifically fine-tuned for financial text and enhances it with data from the **GDELT database**. These scraped articles, enriched with metadata, are cleaned, pre-processed, and stored in the **news_articles.csv** file.

To boost accuracy, the platform employs a **hybrid ensemble model** that combines predictions from FinBERT with those of a **Random Forest Classifier**, ensuring higher reliability by capitalizing on the strengths of both models.

The solution is deployed via **Streamlit** as a user-friendly web application, allowing users to submit articles through text, URLs, images, or audio inputs. It provides real-time feedback on news authenticity, empowering users to make informed decisions. The system also incorporates a **feedback loop**, where user corrections are stored in a CSV file and utilized for periodic model retraining, ensuring continuous improvement.

The project is publicly accessible on **GitHub**, with plans for future enhancements, including transitioning to **time-based retraining schedules**. This robust platform fosters greater confidence in financial markets by equipping users with the tools to effectively detect and avoid fake news.

PROBLEM STATEMENT

As **misinformation** in **financial reporting** arises, **investors** and **markets** face heightened risks. **Manipulative** or **fake financial news** can mislead decision-making and destabilize markets. Traditional detection methods often fall short, missing critical **metadata** and the specialized **language of finance**, resulting in unreliable outcomes. Effective solutions must address these **nuances** to ensure **accuracy** and protect **market integrity**.

PROPOSED SOLUTION OVERVIEW

The proposed solution leverages a **hybrid ensemble model** that combines the strengths of **FinBERT**, a transformer-based model fine-tuned for **financial text classification**, and a **Random Forest Classifier**. While **FinBERT** analyses the **textual content**, the **Random Forest model** processes both **text** and **metadata** from the GDELT database. This ensemble approach enhances prediction accuracy by addressing **complex data patterns** and minimizing the risk of misclassification.

The platform is deployed as a **Streamlit-based web application**, offering users real-time verification of financial news through multiple input modes: **text**, **URLs**, **images**, and **audio files**. It features a **feedback mechanism** that stores user corrections in a CSV file, enabling periodic model retraining to improve performance and adapt to **evolving misinformation trends**.

By hosting the platform on **GitHub**, the solution remains **publicly accessible**, fostering broad adoption and encouraging continuous improvement through **user feedback** and regular **model updates**. This ensures the system stays responsive to changing patterns in **financial misinformation**, empowering users with a reliable tool for **verifying news authenticity**.

SUMMARY OF TECHNOLOGIES USED, FUNCTIONALITIES SUPPORTED, AND ASSUMPTIONS

TECHNOLOGIES USED:

1. MACHINE LEARNING MODELS
    - **FinBERT**: A transformer-based NLP model fine-tuned for financial text classification to detect authentic vs fake news.
    - **Random Forest Classifier**: An ensemble model that processes both text features and metadata to enhance prediction reliability.
    - **Ensemble Approach**: Combines FinBERT's predictions with the Random Forest model for more accurate results.
2. DATA HANDLING AND PREPROCESSING
    - **GDELT Database**: Scrapes news articles and metadata publication date source presence of images.
    - **Pandas & NumPy**: For cleaning preprocessing and structuring data into news_articles. csv.
    - **SMOTE from imbalanced-learn**: Resamples data to address class imbalance and ensure better model training.
3. WEB SCRAPING AND INPUT HANDLING
    - **Newspaper3k**: Extracts article content and metadata from URLs.
    - **Tesseract OCR**: Extracts text from uploaded images for classification.
    - **Google Speech Recognition API**: Converts audio input to text for analysis.
4. FRONTEND AND DEPLOYMENT
    - **Streamlit**: Provides a user-friendly web interface allowing input via text URLs images and audio.
    - **GitHub**: Publicly hosts the web application and code repository to ensure accessibility and transparency.
    - **Matplotlib & Seaborn**: For visualizing model performance accuracy and user feedback trends.

FUNCTIONALITIES SUPPORTED IN THE PROTOTYPE:

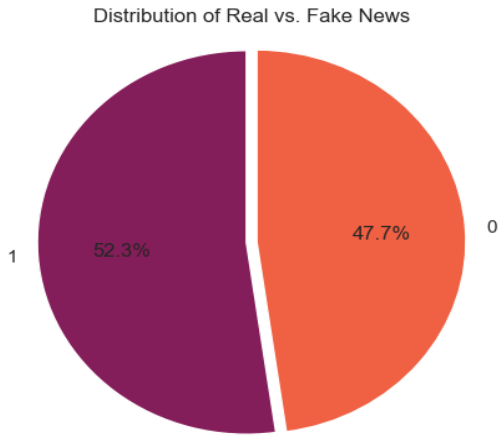
- **Multiple Input Methods**
- **Real-Time Feedback**
- **Ensemble Predictions**
- **Feedback Mechanism**
- **Model Retraining**
- **Visualization and Monitoring**
- **Open Source and Public Access**

Assumptions:

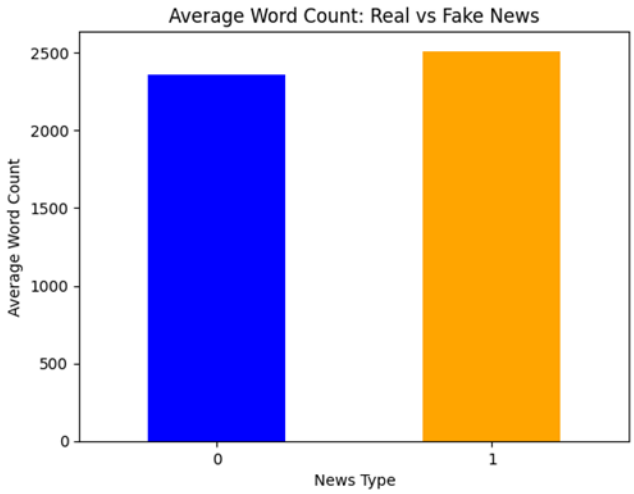
- Representative Dataset: The scraped dataset from the GDELT database accurately reflects real-world financial news patterns and sentiment.
- Consistent Language Trends: The model assumes that the patterns in financial text (e.g., sentiment, terminology) will remain stable over time for reliable predictions.
- Relevant User Input: The system is optimized for finance-related news, and non-financial input may affect prediction accuracy.
- Privacy Compliance: No user data is stored or used without explicit consent. Feedback data is used solely for model improvement and retraining.

VISUAL INSIGHTS

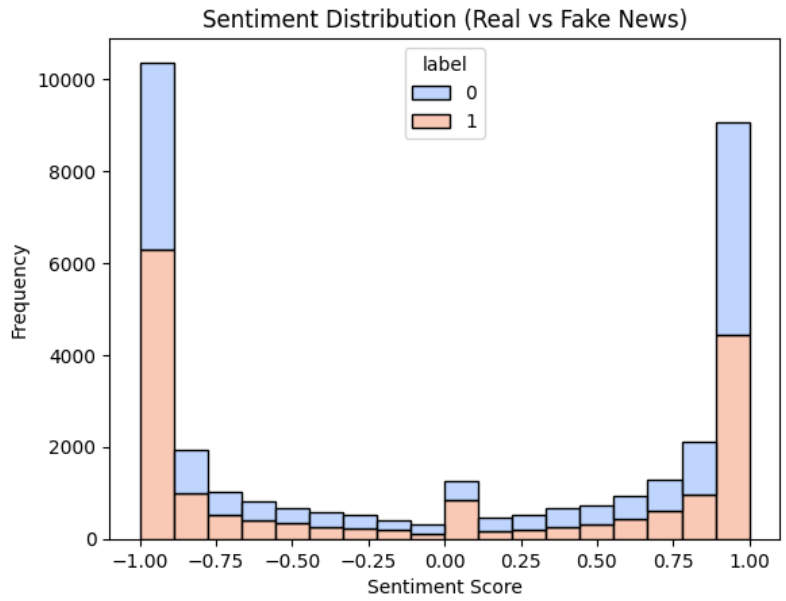
PRIMARY MODEL:



The **Distribution of Real vs. Fake News** shows a well-balanced dataset, with 47.7% fake and 52.3% real news. This balance strengthens the model's learning and prediction accuracy, helping

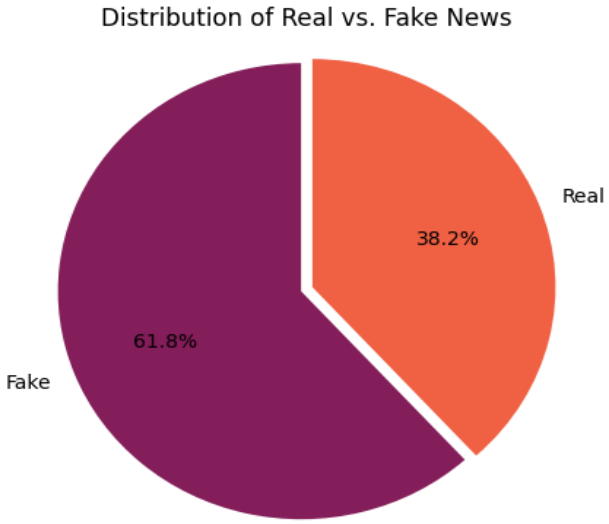


The **Average Word Count: Real vs. Fake News** graph reveals that real news articles are typically longer, offering more detailed content. This helps the model better differentiate between real

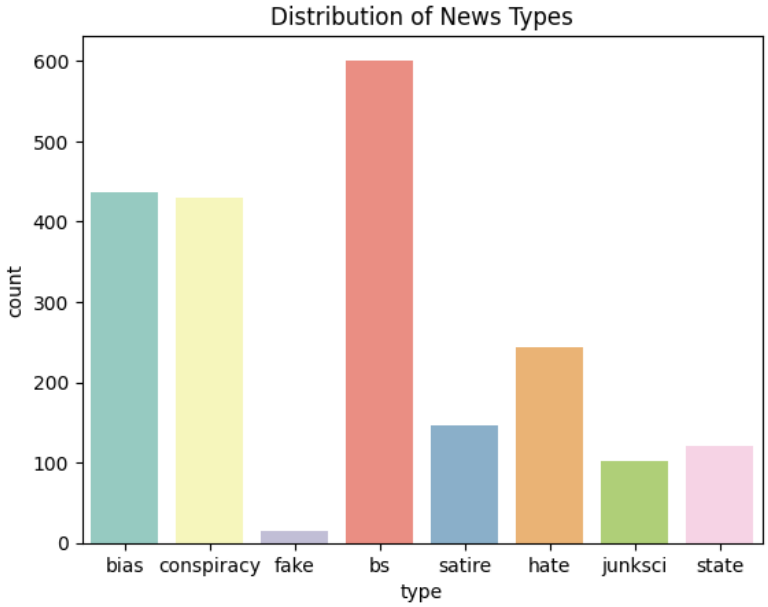


The **Sentiment Distribution (Real vs. Fake News)** reveals that fake news exhibits stronger sentiment extremes, aiding the model in effectively identifying emotionally charged

SECONDARY MODEL



The pie chart shows a higher proportion of fake news, helping the model focus on identifying misleading content and improving its accuracy in detecting fake financial news.

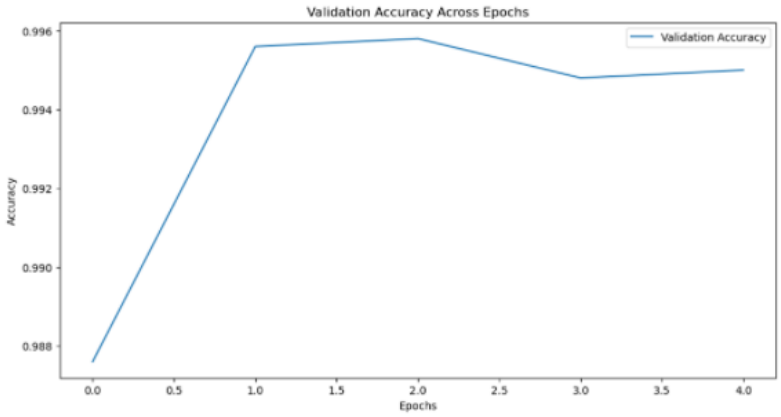


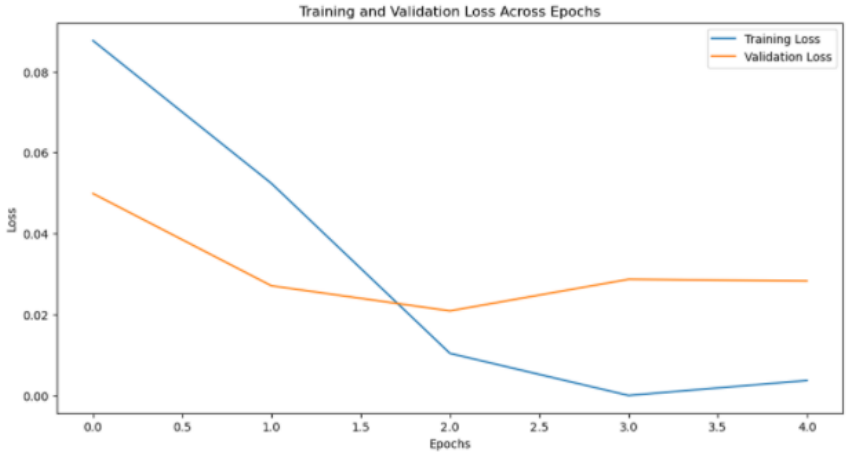
The chart shows the distribution of news types, helping the model better differentiate between various financial news categories for improved accuracy.

ACHIEVED RESULTS: [DEMONSTRATION LINK]

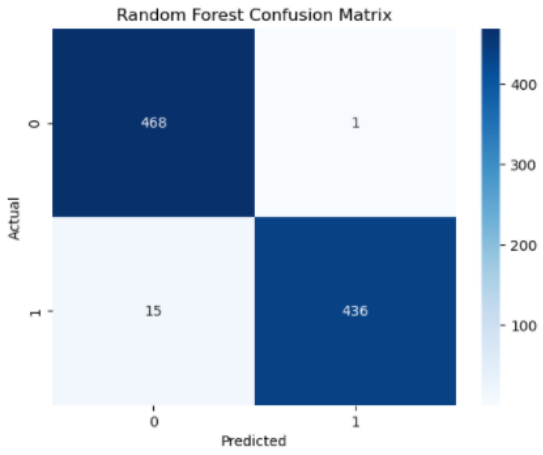
➢ **High Classification Accuracy**

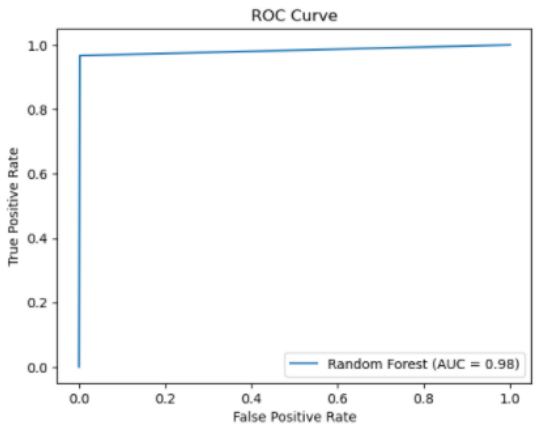
- *FinBERT* achieved an impressive **99% accuracy**, showcasing its strength in classifying financial news text.





- *Random Forest* reached **98% accuracy**, effectively handling text features and metadata.





- The **ensemble model** consistently delivered precision, recall, and F1-scores above **0.98**, minimizing false positives and negatives.

- **Metadata-Enhanced Predictions**
  - The hybrid system combines *FinBERT's text analysis* with *Random Forest's metadata insights* (e.g., publication date, source, image presence), improving predictions even with incomplete or ambiguous text.
- **Real-Time User Interaction**
  - Users can submit news via *text, URLs, images, or audio* through a **Streamlit web interface** and receive **instant classifications** with real-time feedback for a seamless experience.
- **Continuous Improvement via Feedback**
  - Users' feedback on misclassifications is stored in a CSV file, enabling **periodic retraining** that ensures the model adapts to new misinformation trends and stays accurate.
- **Scalable and Open-Source**
  - Hosted on **GitHub**, the platform is **open-source** and scalable, allowing future extensions to other domains like healthcare or political misinformation with minimal adjustments.

CONCLUSION:

The Fake Finance News Detection platform offers a highly accurate, efficient solution for tackling financial misinformation, achieving **99%** accuracy with FinBERT and **98%** with Random Forest. The hybrid ensemble model combines text and metadata analysis for reliable predictions, even in complex cases.

A real-time feedback mechanism ensures continuous model improvement, while public hosting on **GitHub** promotes transparency and collaboration. Designed for investors, journalists, and researchers, this platform strengthens trust in financial markets by minimizing the risks of misinformation.

SCALABILITY:

- **Cross-Domain Expansion:** The system can be adapted to other sectors like healthcare and politics by fine-tuning or replacing FinBERT with specialized models (e.g., BioBERT for healthcare).
- **Multi-Language Support:** It can scale globally by integrating multilingual models like M-BERT or translation APIs, enabling detection across various languages.
- **Cloud-Based Scalability:** Deploying on cloud platforms (AWS, Azure, Google Cloud) ensures the platform can handle high traffic, automate retraining, and remain highly available.
- **API Integration:** The system can provide APIs for seamless integration into financial apps, news aggregators, and corporate platforms, delivering real-time verification.
- **User-Driven Feedback:** A crowdsourced feedback loop will continuously improve accuracy and help the system adapt to new misinformation tactics over time.