AI-Driven Detection of Fake News Beyond Text

Asa Singh 2022113
Sargun Singh Khurana 2022450

Abstract

Fake news has emerged as a critical challenge to societal stability, influencing politics, culture, and online networks. This project focuses on a multi-modal machine learning approach for fake news detection on the Fakeddit dataset. By analyzing both textual and visual features, we trained separate models for text and image data, combined their outputs using an ensemble model, and incorporated metadata to improve classification accuracy. Our work demonstrates the potential of integrating multi-modal data for enhanced fake news detection.
The source code and dataset are available at GitHub Repository.

1. Introduction

The spread of fake news in the digital age has become a critical challenge, manipulating public opinion, eroding trust, and disrupting decision-making. Fake news often combines text with manipulated images on platforms like Reddit, making detection more complex. Traditional models that analyze text or pictures alone fail to capture the interplay between these modalities.

Fake news often involves multi-modal posts where misleading text is paired with images or memes, making it harder for existing models to identify. Additionally, metadata, such as post history and user engagement, should be used more, but it can offer valuable insights into content credibility.

This project proposes a multi-modal machine learning approach integrating text, images, and metadata for more accurate fake news detection. By training separate models for each modality and combining them in an ensemble framework, we aim to improve detection performance and reliability.

The significance of this work lies in its ability to leverage diverse data types—text, images, and metadata—to create a more robust fake news detection system. Using the Fakeddit dataset, tailored for fake news detection, we aim to enhance classification accuracy and offer a more effective solution for combating misinformation in today's digital landscape.

2. Literature Survey

Traditionally, machine learning (ML) models for fake news detection have focused on text-based features. Techniques like TF-IDF (Term Frequency-Inverse Document Frequency) are commonly used to identify essential words in news articles, helping to spot potentially misleading content. However, these models are limited because they analyze text alone and miss important visual clues, such as images and comments.

Error Level Analysis (ELA) has been widely used for image-based detection. ELA detects discrepancies in an image's error levels, highlighting areas that may have been altered or manipulated. This technique effectively identifies fake or AI-generated images but struggles with more complex image manipulations and requires additional methods to enhance accuracy.

While TF-IDF and ELA show promise in their respective domains, they are restricted when used in isolation. Fake news often includes text and images, and analyzing them separately overlooks critical context. This has led to multi-modal approaches, combining textual and visual features for a more robust solution.

By merging text features from TF-IDF with image features from ELA, multi-modal methods offer improved accuracy. The Fakeddit dataset, which includes text, images, and metadata, provides an ideal platform for developing more effective fake news detection models.

3. Dataset and Preprocessing

3.1. Dataset Overview

We use a Reddit dataset with over 1,000,000 entries, originally labeled for 2-way, 3-way, and 6-way classification. Each entry contains a title, image URL, and associated comments, which serve as the core features. For this paper, we focus only on 2-way classification and use a random balanced subset of 60,000 entries, ensuring minimal missing values. The dataset preprocessing involves handling missing data and structuring it for machine learning:

- Entries without associated comments (null linked submission ID) have `num_comments` set to `0`.

- Missing upvote ratio and score are imputed with the median values of the respective fields.
- Missing author and domain fields are replaced with empty strings as placeholders.
- Entries without an image or label are dropped, as they lack essential features for the analysis.

The final dataset contains text, image, and metadata features, with each entry now properly formatted for classification.

3.2. Image Preprocessing

For image data, we apply **Error Level Analysis (ELA)** to detect potential image manipulations by comparing the original image to a recompressed version. The ELA process highlights discrepancies between unaltered and tampered regions of the image. After generating the ELA images, various features are extracted for classification, including:

- **Statistical Features:** Mean, standard deviation, skewness, and kurtosis of pixel intensity distributions.
- **Entropy:** A measure of randomness in pixel intensities.
- **Texture Features:** Extracted using the Gray-Level Co-occurrence Matrix (GLCM), including contrast, correlation, energy, and homogeneity.
- **Edge Features:** Derived using Sobel and Canny edge detectors, capturing edge intensity and edge count.
- **Frequency Features:** High-frequency energy components from the Fourier Transform.

These features are then utilized to differentiate between authentic and tampered regions in the images.

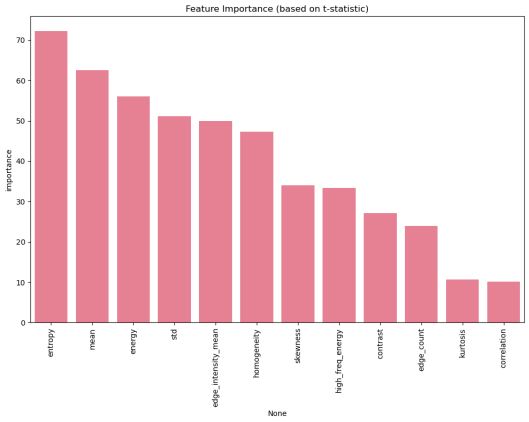

Figure 1. Visualization of the importance of extracted features from ELA images for detecting tampered regions.

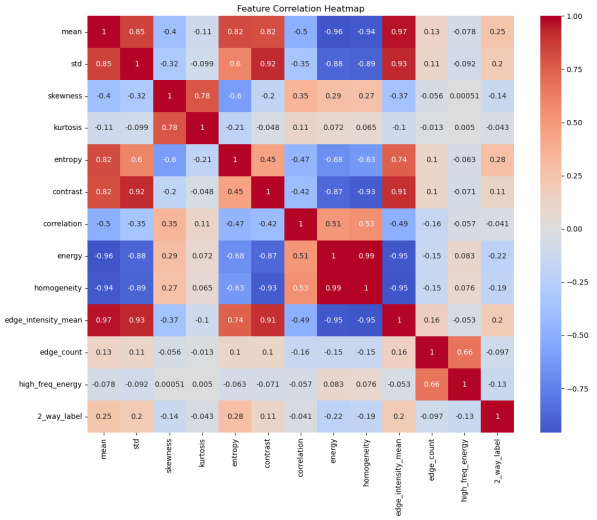

Figure 2. Correlation Heatmap the extracted features from ELA images.

3.3. Text Preprocessing

The preprocessing pipeline consisted of the following steps:

- **Text Cleaning:** Null values in the dataset were removed, followed by cleaning the text by removing URLs, punctuation, and converting text to lowercase.
- **Data Merging:** Titles (`clean_title`) were merged with corresponding comments (`body`) using a dictionary-based lookup, resulting in a new feature `combined_text`.
- **Feature Extraction:**
  - Textual data was transformed into numerical features using the TF-IDF vectorization technique.
  - Stop words were removed during vectorization to reduce noise and focus on relevant terms.

3.4. Preprocessing for Ensemble Model

In the preprocessing step for the ensemble model, predictions from the final text model and the final image model were integrated into the original dataset. These predictions, combined with other metadata features, were used to train the ensemble models. Each entry in the updated dataset contains the following features:

- **Text Features**: Represented using the **TF-IDF** vector.
- **Image Features**: Extracted via **Error Level Analysis (ELA)** and subsequent feature engineering.
- **Metadata Features**: Includes the **upvote ratio**, **score**, **number of comments**, and **year of upload**.

- **Model Predictions**: Probabilities or class predictions from the **final text model** and the **final image model**.

This enriched dataset, combining multi-modal predictions and metadata, serves as input for the ensemble model, which aims to improve the accuracy of fake news detection.

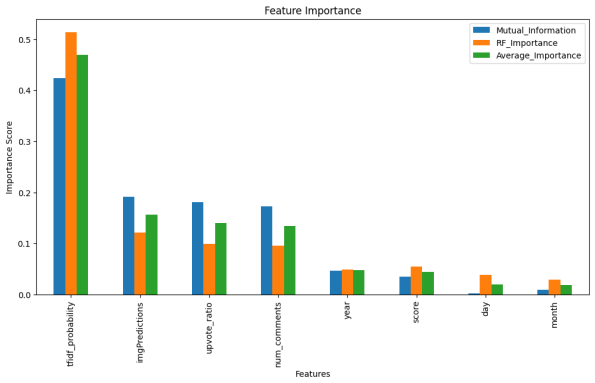

Figure 3. Feature importance for training the ensemble model, including text predictions, image predictions, and metadata.

4. Methodology and Models

4.1. Architecture Overview

Our approach involves three stages:

1. **Image Model:** Trained on ELA features using similar classifiers.

2. **Text Model:** Trained on features from TF-IDF and Bag of Words using classifiers like Random Forest and XGBoost.

3. **Ensemble Model:** Combines predictions from the text and image models along with metadata for the final output.

4.2. Image Model Details

For the image model, **Error Level Analysis (ELA)** was computed for each image to detect manipulations. Additional features, such as edge color, texture, and statistical metrics (e.g., mean, standard deviation), were also extracted. Multiple classifiers, including **Random Forest**, **SVM**, and **XGBoost**, were trained with optimized hyperparameters.

After evaluation, **XGBoost** was selected as the final image model due to its superior accuracy. The predictions from this model were added as a new feature in the dataset to enhance the performance of the ensemble model.

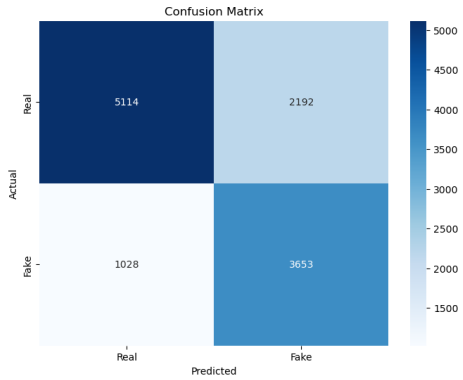

Figure 4. Confusion Matrix for Image-only Model (XGBoost)

4.3. Text Model Details

For the text model, preprocessing steps include removing punctuation and converting all words to lowercase. The text data is vectorized using **TF-IDF** with vector sizes of 2000 and 3000. Several classifiers, including **Logistic Regression**, **SVM**, **Decision Trees**, and **Random Forest**, were trained with optimized hyperparameters.

After evaluation, **Random Forest** was selected as the final text model due to its superior performance. The output probabilities generated by this model were added as a new feature to the dataset to enhance the ensemble model.

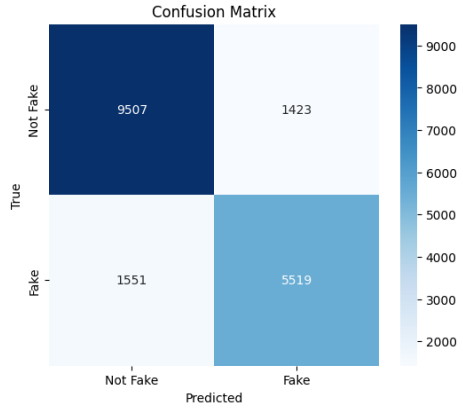

Figure 5. Confusion Matrix for Text-only Model (Random Forest)

4.4. Ensemble Model

The ensemble model integrates the predictions from the text and image models using a meta-classifier, ensuring a comprehensive analysis by incorporating text, image features, and metadata for the final classification.

After evaluating various meta-classifiers, including **Logistic Regression**, **XGBoost**, and **Random Forest**, **XGBoost** was selected as the final ensemble model due to its superior performance. This model aggregates predictions

from the text and image classifiers, along with metadata, to produce the most accurate results.

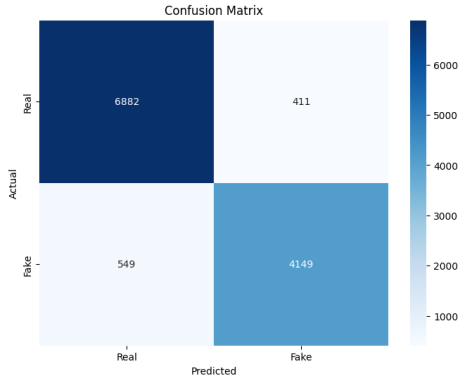

Figure 6. Confusion Matrix for Ensemble Model (XGBoost)

5. Results and Analysis

Our experiments demonstrated that the ensemble model outperformed unimodal approaches in both accuracy and robustness. The results of the individual models and the ensemble model are as follows:

- **Image-only model** (XGBoost) achieved an accuracy of 73.14%.
- **Text-only model** (Random Forest) achieved an accuracy of 83.48%.
- **Ensemble model** (XGBoost) achieved an accuracy of 91.99%.

The significant performance improvement of the ensemble model can be attributed to the combination of predictions from both the text and image models, along with additional metadata, providing a more comprehensive analysis for classification. This multi-modal approach ensured higher accuracy and robustness compared to the unimodal models.

Figures 4, 5, and 6 show the confusion matrices for the image-only, text-only, and ensemble models, respectively.

6. Conclusion

This project demonstrated the power of integrating multiple modalities for effective fake news detection. By combining text, image, and metadata features, we significantly improved the accuracy and robustness of the detection system. The ensemble model, which aggregates the predictions from the text and image models, showcased superior performance over individual unimodal models. The integration of additional metadata played a crucial role in refining the classification results.

However, challenges remained, particularly in terms of preprocessing and handling diverse data types. The text and image data required specialized preprocessing pipelines to ensure that the features could be effectively used by the models. Moreover, optimizing the ensemble model for better performance was a crucial step in achieving the best results.

Future work will explore the application of deep learning techniques to further enhance feature extraction from text and images, as well as improve the overall detection capabilities. Additionally, expanding the dataset and incorporating more sophisticated metadata may further improve the accuracy and generalizability of the model.

6.1. Contributions

- Asa Singh: Dataset cleaning, sampling, and preparation for model training. Development of the image model using Error Level Analysis (ELA) for feature extraction. Creation and integration of the ensemble model combining text and image model predictions. Contribution to report writing and documentation of methodologies and results.
- Sargun Singh Khurana: Text data cleaning, preparation for model training, TF-IDF vectorization for feature extraction, Implementation of Text Model using hyperparameter tuning. Contribution to report writing and documentation of methodologies and results.

References

- N. Krawetz A Picture's Worth... 3 of 31 Copyright 2007 Hacker Factor Solutions, presented at Black Hat Briefings USA 2007. A Picture's Worth... Digital Image Analysis and Forensics
- entitize / Fakeddit
- Feature Extraction based Text Classification: A review May 2022 Journal of Algebraic Statistics 13(1):646-653