Optimizing Fake News Classification Performance: Comparative Analysis using Machine Learning Models with Hyperparameter Tuning

Contribution

Team Member	Contribution
Ayaan	Performed data cleaning and preprocessing, trained all three machine-learning models on both title-based and full-text datasets and writing Results section.
Anubhav Shakya	Conducted error analysis for all three models, including coding the error-analysis procedures and writing the error-analysis section in the report.
Asif	Wrote the Limitations and Ethical Considerations sections, addressing dataset biases, model risks, and responsible use of fake-news detection systems.
Naseeruddeen	Performed exploratory data analysis (EDA), and wrote the Conclusion and References sections for the final report.

Abstract

This report assesses and compares the effectiveness of three commonly used machine learning techniques: Logistic regression (LR), Random Forest (RF), and Naïve Bayes (NB), with an emphasis on hyperparameter tuning to improve their performance. For feature extraction, TF-IDF (Term Frequency-Inverse Document Frequency) was employed, while hyperparameter optimization was carried out using Grid Search CrossValidation (CV). Multiple models are trained based on the title, body and title+body of the article. Out of these models, the Random Forest performed the best across various features, and achieved the accuracy of 92.42% on the title, 98.97% on the body and 99.31% on title+body, outperforming both LR (94.12%, 98.9%, 99.2%,) and NB (93,52%, 95.44%, 95.7%). This report includes the error analysis which revealed the pattern in the misclassified samples. It shows that the shorter and politically neutral articles are more likely to be predicted as false. The findings show how we can use the traditional machine learning techniques that can provide high accuracy in detecting the fake news if the well-engineered text features are applied.

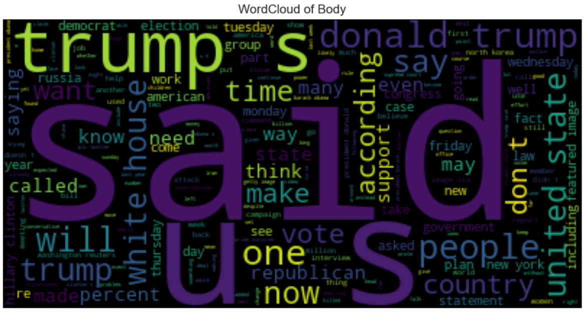
Dataset and Problem

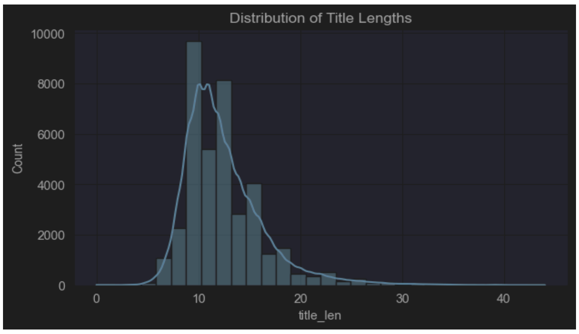
In this study, we used the Fake and Real News dataset from Kaggle that contains 23,502 fake and 21,417 real news articles. Each record in this data has a title, full text, subject category, and publication date. These real and fake news data is in separate files, so we merged both the files to make a complete dataset of 44,898 records. This dataset is balanced enough to use it directly for the training and the next steps without applying any data balancing techniques.

The problem that was created is the binary classification of the data where the goal is to assign the labels to each article, and the labels are: Fake (0) or Real (1). The main metrics that is used to evaluate the models are accuracy, precision, recall, F1-score, and the confusion matrix. We used these metrics because these provide the insights in both overall and class specific performance. The main goal of the project is to assess whether the traditional machine learning techniques are sufficient for this type of classification.

Methods

1. EDA





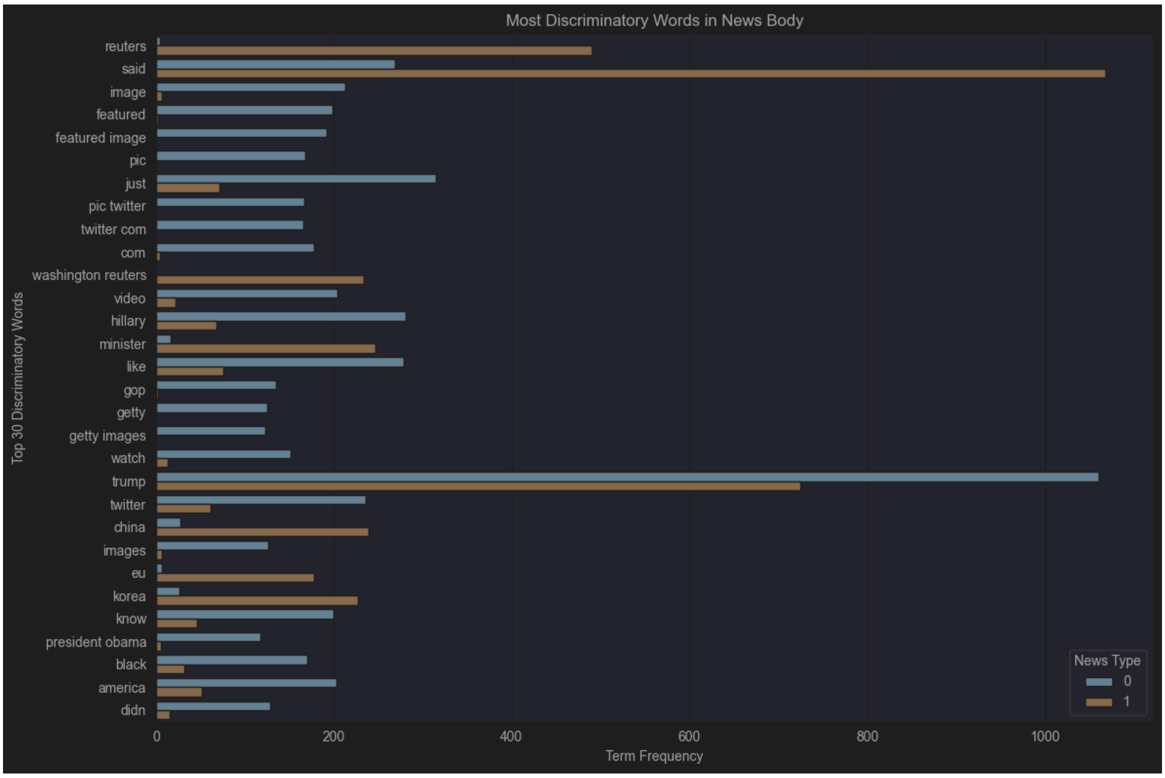


Table.3

2. Preprocessing

The dataset has been cleaned by converting it in lowercase, removing the URLs, special characters, and any extra whitespace character. The English stop words are also removed from the dataset at the state of TF-IDF. This step ensures that the tokens that are useless do not make any influence in the training process of the model. The dataset then split into a train and test set using the 80-20 split to ensure the unbiased evaluation.

3. Feature Extraction

Feature extraction is performed by using the TF-IDF vectorizer. We applied the TF-IDF vectorisation on the dataset up to 50,000 features, which encodes each article from the dataset based on the weighted frequency of its words. The feature space is high dimensional but sparse, which aligns well with the linear models such as Logistic Regression.

4. Baseline Model

The majority class baseline classifier was created on the title and body and it achieved the accuracy of 54.4%. It reflects that the results are slightly imbalanced towards the real news and if we apply any meaningful machine learning models on this, that can outperform this baseline model.

5. Models and Training

These three classical models are used to train on that dataset:

a. Naïve Bayes
b. Logistic Regression
c. Random Forest Classifier

Initially, we trained these models only on the title of the news articles. It shows a good accuracy of around 94% that shows a reasonable performance. However, we trained the models on the full body and both title and body of the news article using the TF-IDF features and all the models improved substantially.

6. Hyperparameter Tuning

We tuned all the 3 models the Logistic Regression, Random Forest and Naïve Bayes by using the GridSearchCV. The best hyperparameters for these are :

Logistic Regression: {'C': 10, 'penalty': 'l2', 'solver': 'liblinear'}

Naïve Bayes: title - {'alpha': 0.5}, body - {'alpha': 0.1}
Random Forest: {'max_depth': None, 'n_estimators': 100}

Results

The Logistic Regression and the Random Forest, both the models achieved the similar and near perfect performance. However, due to its higher accuracy (0.99), higher F1 Score (0.993653), stability across body and combined text, ability to handle nonlinear relationships + feature interactions, we selected the Random Forest as the final model. The Precision and the recall values for both the classes are high, which suggests that the classifier did not favour only one class over another.

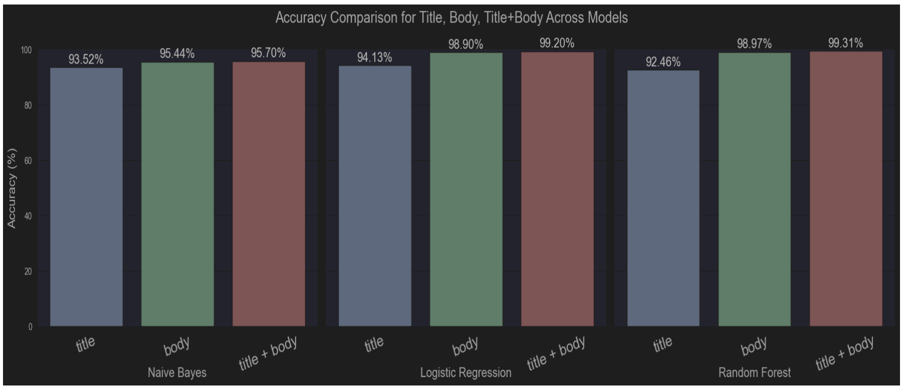


Table:1

Error Analysis

Even though all three models have achieved high accuracy, still these have shown some patterns in the misclassifications. Naïve Bayes performed the weakest in classifying as it misclassified the 334 news articles in both the classes. However, Logistic Regression and Random Forest also made few mistakes. Logistic Regression misclassified 68 news articles while Random Forest 53. There are rare false positives which are fake → real while most of the errors are false negatives where the models have predicted the real news as fake.

In error analysis, we show a consistent trend regarding the length of the news articles. The articles that are misclassified are longer on average but the correctly classified are shorter. It shows that the longer and more complex articles make it hard in the classification. Overall, the Logistic regression showed the most stable and balanced error behaviour and naïve bayes struggled with both the length of articles, shorter and longer. Random Forest produced strong results but still it showed some error in classification.

Limitations:

We knew that its limitations had to be considered, when our system for classifying fake news showed such high accuracy. One of the most common problems implies that the model won't work well with news from other languages, cultural backgrounds or brand-new, more conversational social media posts.

One of the biggest advantages of TF-IDF features, which we utilised in the system, are their effectiveness. Nevertheless, they only care about word frequencies and completely miss the point of what the words mean. Satirical news sites that mimic the style of real news sources can also be classified as genuine.

Another thing that we did not get right was the temporal aspect, basically, news language changes very quickly with new words, events and styles coming to light all over the place. If our model was trained on data from 2016 it would struggle to comprehend articles from 2024 because the language and subjects have completely changed. Which means the system will never be 100% up to date, and will always need to be retrained.

Ethical Considerations

When using automated fake news detection systems, the main ethical concern that arises is the possibility of censoring and suppressing legitimate speech, and this becomes a serious issue if the system is used at a large scale on social media platforms or news aggregators. Authentic journalism could be mislabelled as fake news, and this is something that independent journalists, whistleblowers and people from the less heard communities fear the most, as their writing styles do not fit the general norms. Well-known questions are also raised as to who gets to decide what constitutes fake news.

Our model is trained on a pre-existing dataset, which means someone else made those judgments. News is not always clear-cut, and things like opinion pieces, satires, speculations and stories with thin evidence fall into grey areas. A simple, automated system that slaps a label of fake on all of these would be grossly oversimplifying and could end up stifling speech that is perfectly fine.

Fake news detection is possible, however, and it would be an error to not develop it. What this scenario presents to us is that we must create a thoughtfully implemented strategy with failsafes, human supervision and clear-cut processes and regular checks to monitor the performance of the technology and its influence on society. On its own, tech will not be able to put an end to the spread of fake news, a mix of technical tools, media education, lawmaking and a genuine community commitment to the truth are required to combat it.

Conclusion & Future Work

In a nutshell, a complete analysis of the Kaggle dataset is done using various data visualisation methods and models are implemented on it. The various graphs plotted helps us study and understand the dataset better and draws a lot of insightful conclusions. We are able to implement 3 different classification models using 3 different approaches (title, body and the title or body) and draw the necessary outcomes. We have used Naïve Bayes, Random Forest and Logistic Regression Algorithms for the classification. Future work would involve text + image detection and trying out more advanced techniques like deep learning, ensemble methods beyond Random Forest, word embeddings such as Word2Vec or BERT. These could possibly give us a more complete picture, but would come with the price of being computationally intensive and mind-bogglingly complex. The trade-off between a sophisticated model and something that's easily deployed was not thoroughly investigated in this study. A better approach combining multiple information sources would likely improve both accuracy and robustness.

All in all, the third approach Random Forest had proven to give the best results with an accuracy of 99.8%. The conclusion is that the title and body both play an important role in the classification of news.

References

[1] **Alghamdi, J. (2022).** "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection." *Information*, 13(12), 576. https://doi.org/10.3390/info13120576

[2] **Mallick, C., Mishra, S., & Senapati, M. R. (2023)**. "A cooperative deep learning model for fake news detection in online social networks." *Journal of Ambient Intelligence and Humanized Computing*, 14(4), 4451–4460. https://doi.org/10.1007/s12652-023-04562-4

[3] **Hu, L., Wei, S., Zhao, Z., & Wu, B. (2022)**. "Deep learning for fake news detection: a comprehensive survey." *AI Open*, 3, 133–155. https://doi.org/10.1016/j.aiopen.2022.09.001

[4] **Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2023).** "SEMI-FND: Stacked ensemble based multimodal inferencing framework for faster fake news detection." *Expert Systems with Applications*, 215, 119302. https://doi.org/10.1016/J.ESWA.2022.119302

[5] **Zhang, C., Gupta, A., Qin, X., & Zhou, Y. (2023).** "A computational approach for real-time detection of fake news." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2023.119656

[6]fake-and-real-news-dataset.**https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset?**