Tackling Fake News Classification with Machine Learning

JULY 30

Authored by: Midhun Mathew

I. Introduction

In the age of digital information, fake news - misinformation propagated through social media platforms - poses a significant threat to societal stability and unbiased public sentiment. This report explores leveraging machine learning to classify news articles, providing potential solutions to flag fake news and thereby restore public trust in media platforms.

II. Problem Statement

The widespread proliferation of fake news, often causing bias and social chaos, is of grave concern worldwide. Consequently, the importance of automated detection systems for fake news has become more prevalent than ever. The key areas of value addition are:

- **Misinformation Detection:** Classify fake news by analyzing the content of news articles.
- **Topic Analysis in Fake News:** Automated identification of major themes prevalent in fake news articles.
- **Social Media Moderation:** Enable social media platforms to automatically flag or filter fake news.
- **Building Public Trust in Media:** An informed public that can rely on the veracity of news content is crucial for the healthy functioning of society.

III. Subject Matter

According to a study from the University of Southern California (USC), the propagation of fake news across social media platforms is an inevitable consequence of rewarding users for information sharing. As social media networks grow and usage increases, the spread of fake news is likely to rise. Therefore, it becomes imperative to leverage classification algorithms in machine learning to effectively counter this issue.

IV. Dataset

For this project, data was sourced from a community competition hosted by the University of Tennessee's Machine Learning Club on Kaggle. The dataset includes news articles scraped from various sources, each containing an ID, title, author, text, and binary label (real or fake). The dataset was divided into training and testing CSV files.

V. Data Cleaning and Preprocessing

To ensure a balanced dataset, null values found in the news articles, accounting for less than 1% of the total entries, were deleted. The title and text of news articles were merged into a single column, and redundant columns such as ID and author were removed. Finally, stopwords were eliminated from the news articles and titles.

VI. Exploratory Data Analysis (EDA) Insights

- **Preliminary Analysis**: The initial exploration of the data commenced with a focus on understanding the balance between real and fake news articles. It was reassuring to note that the dataset was well-balanced, mitigating concerns related to model bias towards a particular class during training.
  Further scrutiny of the data involved the analysis of textual characteristics. The average word count and unique word count were compared in both real and fake news articles. It was discovered that real news articles tended to possess a higher word count on average. Additionally, real news articles also showcased a greater number of unique words compared to fake news articles. These findings suggest

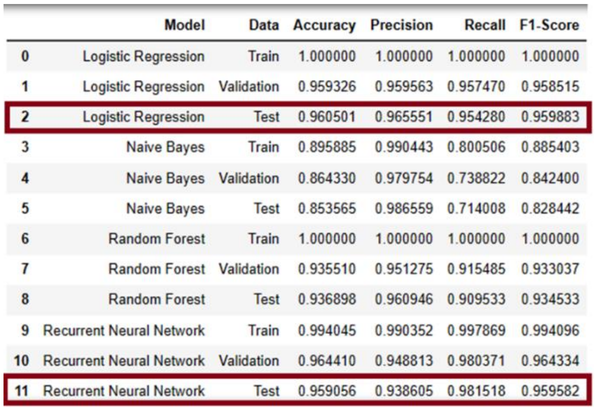
that fake news articles might follow a certain template or set of phrases, leading to lower unique word counts.

- **Data Distribution**: A key insight in the analysis was the higher standard deviation in the average word count of fake news. This observation implies a wider spread in the length of fake news articles, possibly suggesting a lack of consistency or standard in fake news compared to real news. To understand the content distribution of the news articles, the distribution of word counts was explored across real and fake news. Both types of news exhibited a right-skewed distribution. However, a notable difference was the presence of more outliers in the word count for fake news, with most of the real news falling between 0 and 4000 words.
- **Topic Modeling**: Topic modeling, a method to uncover hidden semantic structures in text, was then employed to identify prevalent themes in both real and fake news datasets. The dominant topics seemed to revolve around US Presidential Elections and US Politics. This process of topic modeling could be valuable in recognizing patterns and trends in fake news, providing an additional feature for our machine learning models.
- **Hypothesis Testing**: Finally, hypothesis testing was conducted, including Two-sample t-tests and Chi-square tests of independence, on the real and fake news datasets. The tests showed a significant difference in the average word count between real and fake news articles, confirming our initial analysis. Furthermore, the Chi-square test indicated a significant difference in the occurrence of common words between real and fake news, underscoring the potential importance of these word frequencies as features in our classification models.

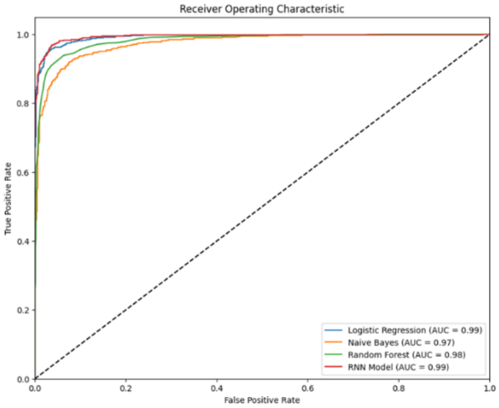
VII. Data Modelling and Results

This study evaluated four different machine learning models: Logistic Regression, Naive Bayes, Random Forest, and Recurrent Neural Network (RNN). Each model was judged on its performance on binary classification tasks using metrics such as Accuracy, Precision, Recall, and F1-Score.

- **Logistic Regression**: The Logistic Regression model performed excellently on the training data with near-perfect scores. On the validation and testing data, it maintained high levels of performance, achieving over 0.95 on all metrics. This demonstrates that the model generalizes well to unseen data and effectively balances precision and recall.
- **Naive Bayes**: While achieving high precision, the Naive Bayes model had a lower recall and F1-scores, indicating a higher rate of false negatives. This suggests that the model's assumptions may not fully apply to our data.
- **Random Forest**: The Random Forest model achieved near-perfect scores on the training data and over 0.90 on all metrics on the validation and testing data, indicating a solid overall performance.
- **Recurrent Neural Network (RNN)**: The RNN model performed exceptionally well on all metrics, achieving the highest recall among all models on the testing data, suggesting its strong ability to minimize false negatives.



	Model	Data	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	Train	1.000000	1.000000	1.000000	1.000000
1	Logistic Regression	Validation	0.959326	0.959563	0.957470	0.958515
2	Logistic Regression	Test	0.960501	0.965551	0.954280	0.959883
3	Naive Bayes	Train	0.895885	0.990443	0.800506	0.885403
4	Naive Bayes	Validation	0.864330	0.979754	0.738822	0.842400
5	Naive Bayes	Test	0.853565	0.986559	0.714008	0.828442
6	Random Forest	Train	1.000000	1.000000	1.000000	1.000000
7	Random Forest	Validation	0.935510	0.951275	0.915485	0.933037
8	Random Forest	Test	0.936898	0.960946	0.909533	0.934533
9	Recurrent Neural Network	Train	0.994045	0.990352	0.997869	0.994096
10	Recurrent Neural Network	Validation	0.964410	0.948813	0.980371	0.964334
11	Recurrent Neural Network	Test	0.959056	0.938605	0.981518	0.959582

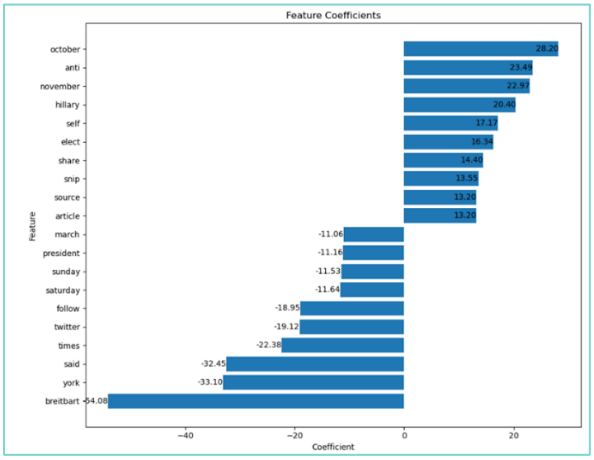


The RNN model showed the highest recall on the testing data, indicating its exceptional ability to minimize false negatives. Logistic Regression and Random Forest models, while simpler, also demonstrated excellent performance, indicating a balance between precision and recall.

VIII. Model Interpretation

Further model interpretation study was conducted on the Logistic Regression model, with an analysis revealing the top 20 words contributing significantly to the model's predictability.

Top Words (Real/Fake)	**Importance Scores**
October	28.20
Anti	23.49
November	22.97
Hillary	20.40
Self	17.17
Breitbart	-54.08
York	-33.10
Said	-32.45
Times	-22.38
Twitter	-19.12



IX. Conclusion

While all models demonstrated commendable performance, the RNN model stood out due to its balance of precision and recall, particularly on the test data. The report suggests that for the task at hand, utilizing temporal information with RNNs can yield better performance. However, model selection should also consider factors like interpretability and computational efficiency, where simpler models like Logistic Regression and Random Forest may be preferred depending on specific task requirements. Another key observation is as the model was trained majorly on political news, it showed high bias to political news topics. Hence, this model may not generalize well and not offer high accuracy on non political news topics. However, on the positive side, the high accuracy of the model is promising and by training news articles from other topics model bias could be improved in the future.