Flip Robo Technologies

Fake News Project

Submitted by: John Tojo

Data Science Intern at Flip Robo Technologies

ACKNOWLEDGMENT

It gives me immense pleasure to deliver this report. Working on this project was a great learning experience that helped me attain in-depth knowledge on data analysis process. Flip Robo Technologies (Bangalore) provided all of the necessary information and datasets, required for the completion of the project. I express my gratitude to my SME, **Gulshana Chaudhary**, for providing the dataset and directions for carrying out the case study procedure.

INTROUDCTION

The advent of the Internet and social media has led to an explosion in the spread of information. While this can be a positive development, it also makes it easier for misinformation or fake news to spread. Fake news is deliberately misleading or false information designed to mislead people. It has become a major problem as it affects individuals and society as a whole, creating misconceptions and mistrust. Fake news can be fuelled by a number of factors, including political agendas, propaganda and economic interests. In some cases, the media may spread fake news to attract viewers and generate online advertising revenue. Therefore, it is very important to detect fake news and prevent it from spreading further.

Given the impact of fake news on individuals and communities, it is important to have a way to detect it. Developing tools to detect fake news is essential to combating disinformation. These tools can help verify the accuracy of news and prevent the spread of misinformation. In conclusion, fake news is on the rise and demands our attention and action. Through better vetting processes, education and the development of fake news detection tools, we can reduce the impact of fake news on our lives and ensure that accurate information is shared.

Problem Statement

The spread of fake news is becoming a major problem in today's society. With the advent of social media, fake news is easily shared and spread among people, leading to widespread misinformation. Fake news often sets political agendas and can cause harm to individuals, communities and society as a whole. It can also damage the credibility and reputation of media organizations. Given the negative impact of fake news, there is an urgent need to detect and prevent its spread. The purpose of this study is to address the problem of fake news and develop a method to detect fake news in order to protect people from being misled and maintain the integrity of the media.

DATA SOURCES AND THEIR FORMATS

This study is based on a dataset in CSV format collected from FlipRobo. The dataset consists of two separate files, one containing real news stories (21417) and the other containing fake news (23481). The data in the file is organized into columns, including title, text, subject, and date. The Title column indicates the name of the news report, the Text column provides a detailed description of the headline, the Subject column indicates the type of news involved in the report, and the Date column indicates the date of publication of the news. This dataset helps to understand the characteristics of real and fake news and to detect fake news.

Exploratory Data Analysis (EDA)

To prepare the data for further analysis, a new column named "Target" was added to the data frame that distinguished between true and fake news. The values in this column were filled with "Fake" for fake news and "True" for true news. The data frame was then encoded using "get_dummies" where "Fake" was given a value of 1 and "True" was given a value of 0, and stored in the "Label" column.

The data frame was checked for any null values and duplicates, and it was found that there were no null values present and 209 duplicates which were removed. The "Title" column was selected as the text data, and the texts present in the "Text" column were processed and transformed. The process involved lowercasing the texts, tokenizing the words, removing non-alphanumeric characters, stopwords, and punctuation, performing stemming of words, and joining the processed words into a single string. The transformed data was stored back in the "transformed_title" column.

Visualization

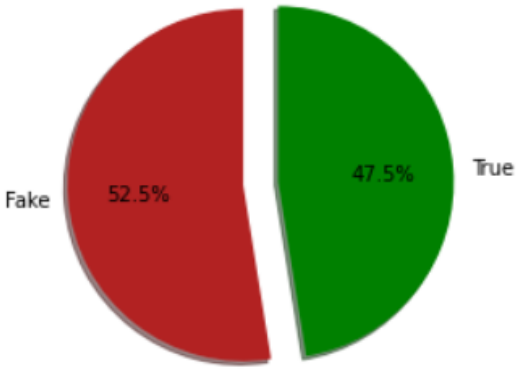


Fig1: pie chart showing the distribution of fake and true news present in the dataset
it can be seen that data set provided is balanced

Fig2: number of texts vs num_words

it can be seen most of the texts have 15 words and most of the texts are contained within 20 words

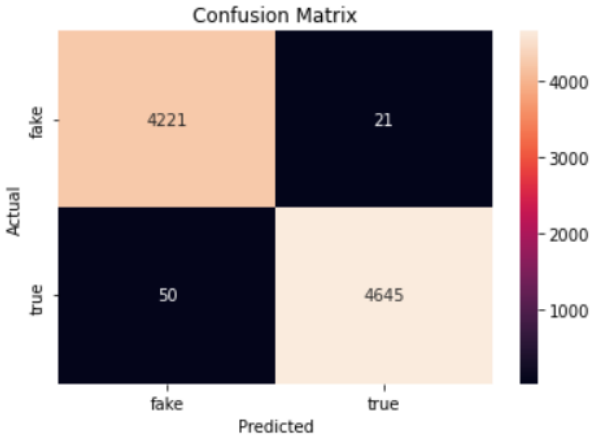
Fig3: showing word cloud for true news

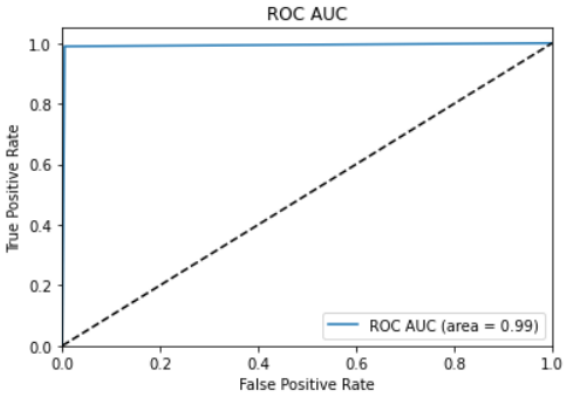
Fig4: showing word cloud for fake news

Model Creation:

The "title" column was selected as the text data, and was lowercased, tokenized, stemmed, and joined back into a single string. The transformed text was then vectorized using one-hot encoding with a vector length of 5000. The target labels were also stored, and both the transformed text and labels were converted to arrays. A sequential model was created with an Embedding layer, 2 Bidirectional LSTM layers, 2 Dense layers and 2 Dropout layers. The model was compiled with binary_crossentropy loss, adam optimizer and accuracy as evaluation metric. The model was then evaluated using 5-fold cross-validation with early stopping as a callback to prevent overfitting. The final result was the mean accuracy of the 5 folds with its standard deviation.

The model includes two bidirectional LSTM layers, each with 100 units, to capture both past and future context information in the input sequences. The first LSTM layer processes the embedded sequence of words in both forward and backward directions, encoding complex relationships between words in the input sequence. The second LSTM layer further processes the hidden states generated by the first layer, applying bidirectional processing to capture additional context information. The final prediction is made using the combined context information from both LSTM layers.





The model was able to achieve

- overall accuracy of 99.21%
- roc_auc of 99.22%
- f1-score of 99.24%
- precision of 99.55%.

Conclusion

The model created for identifying fake news in a dataset showed promising results in terms of accuracy, precision, and f1-score. The 5-fold cross-validation approach used in training and testing the model achieved an overall accuracy of 99.21%, a roc_auc of 99.22%, a f1-score of 99.24%, and a precision of 99.55%. The preprocessing of the transformed title column and the use of an Embedding layer, bidirectional LSTM layers, dropout layers, and dense layers with a sigmoid activation function all contributed to the successful performance of the model.