Gabriel Pesce - Małgorzata Gierdewicz - Tidiane Bengriche

Exploratory Analysis report

In the following report, we will show the results of the exploratory analysis of two datasets:

Dataset 1: Fake News Dataset Detection

The selected dataset contains news articles on various topics such as politics (mostly) and world events.

Dataset information

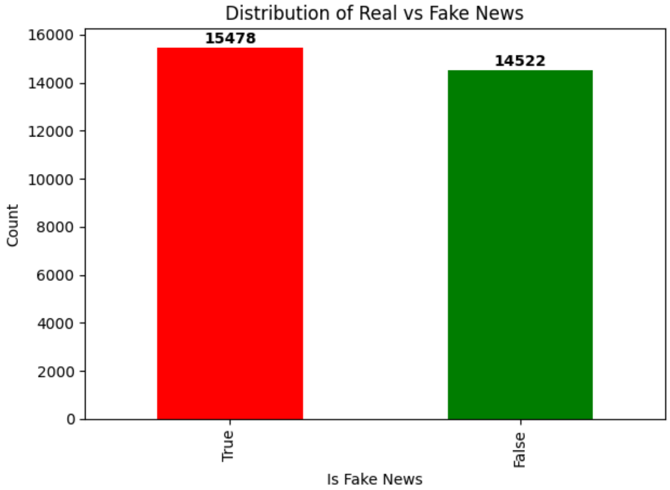
The dataset contains **30k news articles**, distributed across **six columns**. Each entry provides an identifier, title, content, subject, publication date, and a binary label indicating whether the news is **real (1)** or **fake (0)**.

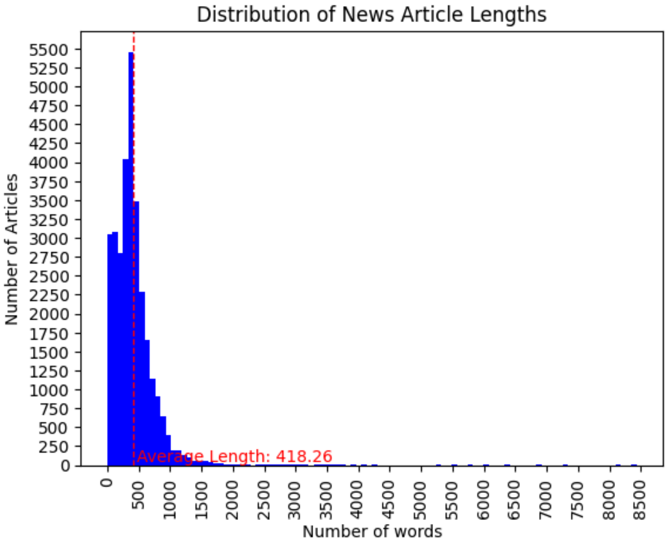
Dataset analysis

During our exploratory analysis, we provided some statistics information:

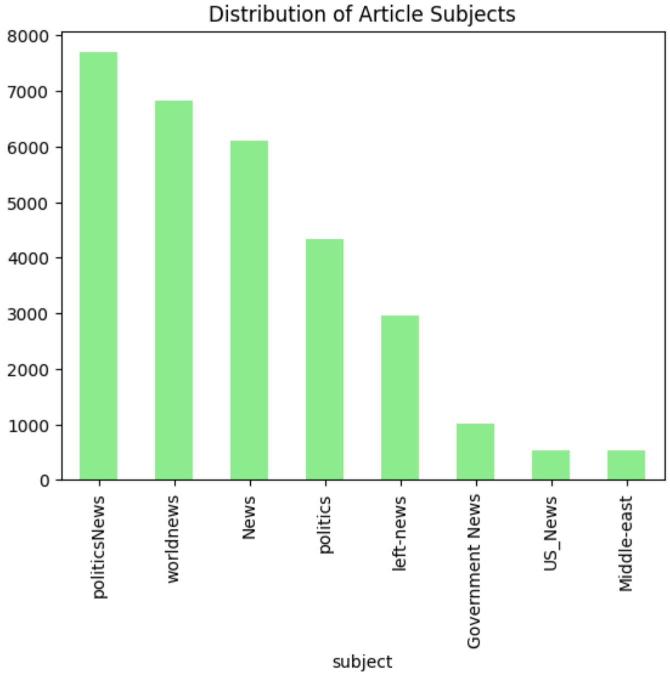
1. We compared the number of Real and Fake news, showing that the dataset is quite balanced.

2. We showed the average length of the articles, showing that the average length of the articles is around **418 words** with the longest article with **8435 words** and the shortest with **one word**.





3. We checked for duplicate rows, finding that there are 92 duplicates.
4. The dataset doesn't have missing or null values.
5. We checked the topic most covered in the articles, resulting in policy news.



Dataset 2: Fake News

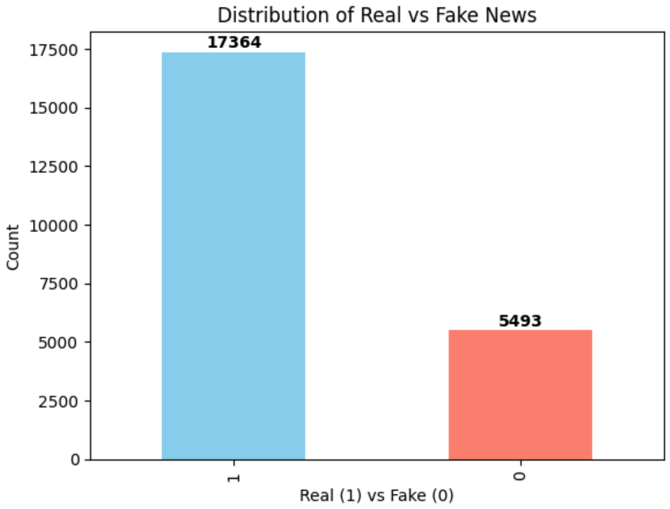
Dataset information

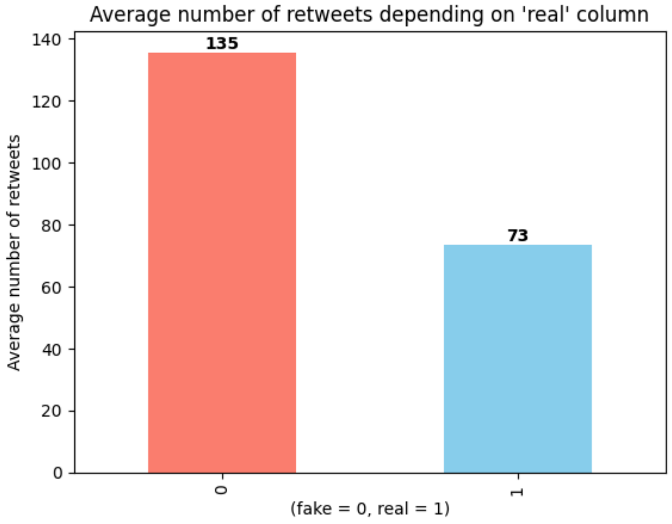
The dataset originally contained **23k news articles**, distributed across **five** columns. Each entry provides a title, link to the news, source domain, number of retweets, and a binary label indicating whether the news is **real(1)** or **fake (0)**.

Dataset analysis

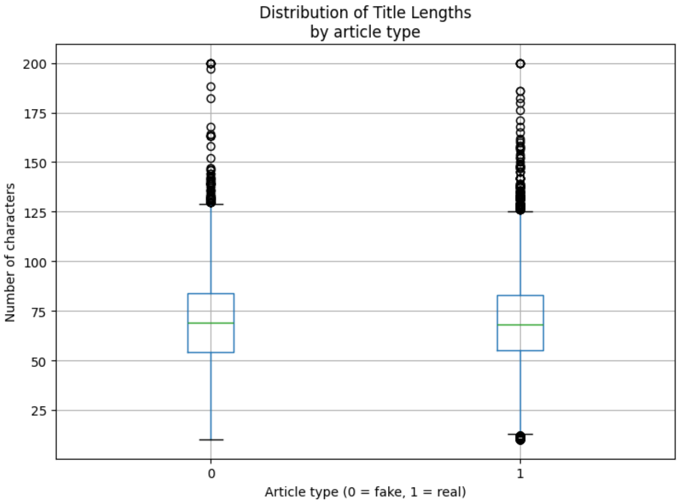
As with the previous datasets, we provided some statistics information:

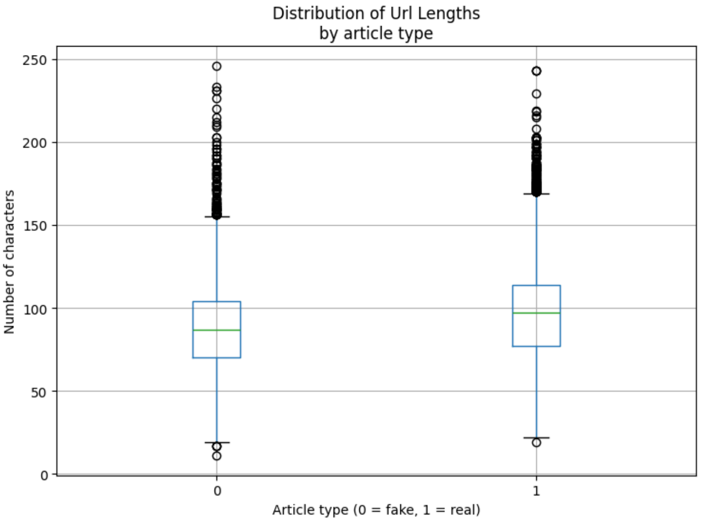
1. We compared the number of Real and Fake news, showing that the dataset is not balanced.
2. We compared the number of retweets between real and fake news, showing that is higher for the real ones





3. We showed the difference in the distributions of title length between fake (0) and true (1) articles using box plots, showing that both types have the same title lengths.
4. We showed the difference in the distributions of url length between fake (0) and true (1) articles using box plots, showing that real ones have longer.





5. We checked for null values, finding 330 articles without news_url and source_domain.
6. We checked for duplicate rows, finding 136 duplicate articles.

Summary and conclusion

Both datasets were analyzed in order to understand the structure of them and calculate some statistical data. The next step will be to process and resize both datasets in order to use the most useful articles as prompts for the chatbots. More information about the analysis (including some graphs and plots) can be found in the notebooks folder in our GitHub repository.