Gabriel Pesce - Małgorzata Gierdewicz - Tidiane Bengriche

Exploratory Analysis report

Dataset 1: Fake News Dataset Detection

The selected dataset contains news articles on various topics such as politics (mostly) and world events. We chose this dataset because it fits the purpose of our methodology: **evaluating the reliability of chatbot responses** when provided with real and fake news content.

Dataset information

The dataset originally contained **30k news articles**, distributed across **five columns**. Each entry provides a title, content, subject, publication date, and a binary label indicating whether the news is **true (1)** or **fake (0)**.

Dataset analysis & refactoring

During our exploratory analysis, we identified several issues that could compromise the reliability of our evaluation:

1. **Duplicate records:**
   Several rows contained identical content under different IDs.
   ➜ These duplicates were removed, keeping only one representative entry per group.

2. **Incomplete articles:**
   Some records contained extremely short or empty content (e.g., one or two words).
   ➜ These were filtered out to ensure meaningful input for chatbot evaluation.

3. **Clearly fabricated articles:**
   Many fake articles used *“clickbait”* or *“shouty”* titles written entirely in uppercase (e.g., “YOU WON’T BELIEVE WHAT HAPPENED…”).
   ➜ Articles with these characteristics were excluded unless they mentioned official entities (e.g., *NASA*, *FBI*, *COVID*) to avoid bias.

Number of news (%)	Before cleaning	After cleaning
Real news	48.75%	68.89%
Fake news	51.25%	31.11%

This shift suggests that a significant portion of fake articles followed an exaggerated or manipulative style.

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 **true (1)** or **fake (0)**.

Dataset refactoring & analysis

As with the previous datasets, some cleaning and analysis operations will be performed :

1. **Incomplete and duplicate articles :**
   In this dataset, some rows are duplicated and contain missing values; therefore, we make sure to remove them to avoid inaccurate results later on.

2. **Check the distribution :**
   In this dataset, 76.3% of the articles are true and 23.7% are false, which creates an imbalance but will not impact the sequence of requests to the chatbot. Moreover, we observe that the number of retweets is, on average, higher for real news than for fake news.

Idea to link the project with Knowledge Graphs

In order to connect our project to knowledge graphs, here is an idea that we would like to implement :

1. **Create the initial Knowledge Graph :**
   The first step is to create a knowledge graph from our dataset, containing both real and fake news.

2. **Make requests to the chatbots :**
   The next step is to send queries to the chatbots using the information from our knowledge graph in order to remove fake news from our datasets.

3. **Create a new Knowledge Graph :**
   The final step is to create the actual knowledge graph using the information returned by the chatbots, containing only genuine news or corrected news.

Summary and conclusion

Both datasets were cleaned and analyzed to ensure high-quality, non-redundant information suitable for evaluating chatbot reasoning and truth verification. The next step will involve using these datasets to test cyclic consistency between chatbot outputs and structured facts in knowledge graphs. More information about the analysis (including some graphs) can be found in the notebooks folder in our GitHub repository.