

Projet:
Creating an Arabic Fake News Detection
Dataset

Réaliser par:

Djeghdjogh Sara

M'ziane Houda

Makeri Atika

Introduction:

The objective of this project is to create a comprehensive dataset for detecting fake news in Arabic. Beyond merely labeling articles as fake or real, the dataset will also categorize them by topic, providing insights into the spread of misinformation across various domains. This report outlines the methodology, data collection process, annotation procedure, and future directions for this project.

Details about the project:

We have collected the data from 3 sites:

- No Rumors (<http://norumors.net/>)
- Verify-Sy (<https://verify-sy.com/>)
- Fatabyyano (<https://fatabyyano.net/>)

We didn't use the first site (<https://misbar.com/>) because we couldn't extract information from it.

For the 2 site: No Rumors. There is a problem in it.so we have scrapping just from the home page.



Script Functionalities:

scrap_norumors_page(): This function retrieves data from the Norumors homepage by extracting the titles and domains of the articles.

scrap_fatabyyano_pages(): This function traverses multiple pages of Fatabyyano to extract the titles of the articles along with their types of news.

extract_data_from_additional_links(): This function iterates through a list of additional links to extract articles and their types of news from the Verify-Sy website.

main (): The main function of the script that calls all the data retrieval functions, combines the articles from all sources, and then writes them into a CSV file named articles.csv.

Analyse du dataset:

1. Le nombre total des lignes est : 3015.
2. Number of articles in each domain:

Religious, Political: 3 articles.

[Unspecified Domain]: 3003 articles.

Political: 6 articles.

Religious: 1 article.

Social, Political: 1 article.

3. Number of articles with each type of news:

Rumor: 12 articles - These are likely false news or rumors.

Misleading: 1268 articles - These articles may contain misleading or biased information.

False: 778 articles - These articles are false or contain untrue information.

[Type of news unspecified]: 708 articles - It may be useful to check these articles to determine their nature.

Fabricated: 2 articles - These articles are fabricated or invented.

Partially False: 135 articles - These articles contain elements of truth mixed with false information.

True: 46 articles - These articles are verified to be true.

Satirical: 27 articles - These articles are humorous or satirical.

Opinion: 38 articles - These articles likely represent opinions or analyses rather than raw facts.

Conclusion:

The Arabic Fake News Detection Dataset, compiled from prominent Arabic news sources, offers critical insights into the prevalence of misinformation. This collaborative effort lays the groundwork for future research and algorithm

development aimed at combatting false information in Arabic-speaking communities.