Title: Creating an Arabic Fake News Detection Dataset

Introduction:

In our project, we aimed to create an Arabic Fake News Detection Dataset by scraping articles from various Arabic websites and categorizing them based on their content. The process involved scraping articles from Misbar, No Rumors, Verify-Sy, and Fatabyyano websites. We then identified whether each article was real or fake and categorized them into different topics.

1. Scraping Articles

Misbar Website :

Firstly, we imported the necessary modules. Among them, there's csv, which enables us to manipulate CSV files, and webdriver from Selenium, utilized for automating the Chrome browser. Additionally, we used BeautifulSoup for parsing the HTML content of the website.

Subsequently, we automated the process of opening a Chrome browser and loading the desired web page. Once the page was fully loaded, we extracted its HTML content and closed the browser.

By utilizing BeautifulSoup, we parsed the extracted HTML content to find links to editorial and fact-checking articles. For each found link, we retrieved the article's URL and its title.

We stored this information in a list. Finally, we wrote this list to a CSV file named articles.csv, where each line represents an article with its associated information.

No Rumors , Verify-Sy and Fatabyyano Website :

We start by defining a list of URLs representing different websites. For each URL, our program retrieves the webpage content using the requests library and parses it with BeautifulSoup. We then search the parsed HTML for article titles, typically located within <h2> tags. Once we've identified a title, we look for the associated URL, often found as the href attribute of the previous <a> tag.

After collecting the article titles and URLs, we append this data to a CSV file

2. Identifying Fake and Real Articles

We've used Python alongside some additional libraries to collect data to identifying Fake and Real Verify-Sy articles. This website offers two types of pages: one confirms news, while the other exposes fake news.

To gather this information, we defined the structure of the URLs for each type of page and decided on the number of pages we wanted to extract data from. Then, we created a function to extract relevant data such as article titles, URLs, and types.

3. Identifying Articles by Category

We have a code snippet that categorizes articles stored in a CSV file based on their titles. The categorization is done using predefined categories with associated Arabic keywords.

We have a dictionary that maps each category to a list of relevant keywords in Arabic.

We have a function that takes an article title as input and matches it against the keywords in each category. If a match is found, the article is assigned to that category; otherwise, it's labeled as 'Other'.

4. statistics

We can generate statistics for the entire dataset by analyzing the categorized articles in the CSV file. These statistics can provide insights into the distribution of articles

- total number of articles : 183
- 39 article from Misbar
- 47 article from no Rumors
- 56 article from verify sy
  50% of the articles are categorized as "Reel".
  50% of the articles are categorized as "Fake".
- 41 article from Fattabayano

Conclusion

In conclusion, our project involved scraping articles from Arabic websites, identifying real and fake news, categorizing articles by topic, and generating statistics on the dataset. This dataset serves as a valuable resource for research in Arabic fake news detection and contributes to efforts in combating misinformation.