Creating an Arabic Fake News Detection Dataset

This report outlines the objectives and methodology for a project aimed at creating a robust dataset for detecting fake news in the Arabic language. The key goals include gathering a balanced mix of authentic and fabricated news articles from reputable fact-checking websites, categorizing them by topic domains such as politics, religion, and health, and establishing a rigorous fact-checking process to ensure the dataset's reliability.

made by :

-Amrouche Nassiba

-Mohamed mahmoud Ikram

Master 1 IA

Project Objectives

The primary objective of this project is to construct a comprehensive dataset of Arabic news articles that can be used to train and evaluate models for detecting fake news. The dataset will consist of an equal number of authentic and fabricated articles, ensuring a balanced representation of real and false information. By categorizing the articles by topic, the dataset will also enable the exploration of domain-specific patterns and biases in fake news propagation.

Another key goal is to establish a reliable fact-checking methodology that can be applied to the article collection process. This will involve cross-referencing the articles against reputable fact-checking sources and expert reviews to verify their authenticity, ensuring the integrity and trustworthiness of the dataset.

Article Categorization

Each article in the dataset will be carefully categorized by topic to enable in-depth analysis and exploration of domain-specific patterns in fake news.

Politics

Articles related to government, elections, public policy, and other political topics will be categorized under the "Politics" domain.

Religion

Articles discussing religious beliefs, practices, and related issues will be categorized under the "Religion" domain.

Health

Articles covering health-related topics, such as medical treatments, disease outbreaks, and wellness trends, will be categorized under the "Health" domain.

Social

Articles that discuss social issues, trends, events, or phenomena that impact society at large.

Fact-checking Methodology

Ensuring the accuracy and reliability of the dataset is a crucial aspect of this project. we implement a rigorous fact-checking process to verify the authenticity of each news article, drawing on multiple reputable sources and expert reviews.

1 Source Verification

The first step in the fact-checking process will be to thoroughly vet the sources of the news articles, ensuring they are reputable and reliable. This will involve cross-referencing the articles against known fact-checking websites and expert assessments.

2 Content Validation

Each article's content will be carefully analyzed and compared against multiple reliable sources to verify the accuracy of the claims and information presented.

Methodologie

1 find_news_articles :

- This function sends a GET request to 'http://norumors.net' to retrieve the HTML content of the webpage.
- It then uses BeautifulSoup to parse the HTML content and extract the relevant information, such as article titles and categories.
- The extracted data is written to the CSV file 'combined_articles.csv' with the specified fieldnames in Arabic.
- in conclusion, this function sends a GET request to 'http://norumors.net' to retrieve the HTML content of the webpage.
- in conclusion , this function extracts articles from the website `http://norumors.net` using BeautifulSoup with the `requests` library.

2 find_news_articles2 :

- Similar to the first function, this one sends a GET request to 'https://fatabyyano.net/' to retrieve the HTML content.
- It parses the HTML content using BeautifulSoup and extracts the article titles, categories, and publication dates.
- The extracted data is appended to the existing CSV file 'combined_articles.csv'.
- in conclusion , this function extracts articles from the website https://fatabyyano.net/ using BeautifulSoup with the requests library.

find_news_articles3 :

- This function uses Selenium WebDriver to automate the web browsing process for 'https://verify-sy.com/'.
- It waits for specific elements of interest to load on the webpage using WebDriverWait from Selenium.
- Once the elements are loaded, it extracts article titles, categories, and publication dates using BeautifulSoup.
- The extracted data is appended to the existing CSV file 'combined_articles.csv'.

Main Function:

- In the main block, all three scraping functions (`find_news_articles`, `find_news_articles2`, and `find_news_articles3`) are called sequentially.
- After scraping from all websites is completed, a message indicating the completion of the scraping process is printed.

statistics about the dataset

Total Number of Articles:

40 articles

Total Number of articles with Complete Information: 25

Distribution Across Topics:

- **Categories**:
  - 7 :دينية articles
  - 18 :سياسي articles
  - 3 :اجتماعية article
- **Health Status**:
  - 12 :إشاعة articles
  - 5 :زائف article
  - 8 :مضلل articles

Publication Dates:

- **Publication Dates Range**:
  - From 2020-08-28 to 2024-04-12
- **Frequency of Publication Dates**:
  - 2024: 13 articles
  - 2023: 2 article
  - 2020: 2 articles

columns :

'المقال', 'التصنيفات', 'صحة الخبر', 'تاريخ النشر'

Insights:

- The majority of the articles are categorized as political, with a significant portion falling under the religious category.
- Nearly all articles are labeled as rumors (إشاعة), indicating a prevalence of misinformation.
- The dataset contains articles published over a span of several years, with the most recent articles dated in 2024.

Conclusion

This project represents a significant step forward in the fight against fake news in the Arabic-speaking world. By constructing a robust and reliable dataset of authentic and fabricated news articles, according to the workflow process:

Data Collection, Data Writing, Error Handling.