SHYAKA NOBLE DAVID ICHERAK BENAMARA MESSAR AYA

Arabic Fake News Detection Dataset

Introduction

Fake news detection is a critical task in today's digital age, particularly with the rise of misinformation on social media platforms. In this report, we elucidate the methodologies employed to create an Arabic dataset for fake news detection. We also provide pertinent statistics about the dataset, such as the total number of articles and their distribution across topics.

Data Collection

Two main methodologies were employed for data collection:

Scraping with BeautifulSoup

Used to get view the structure of the html and locate tags. BeautifulSoup, however, could not be used any further as it has certain limitations like working around dynamic elements like buttons.

Scraping with Selenium

The dataset was therefore collected using Selenium, a tool commonly used for automating web browsers. Selenium was used to navigate the Misbar website, scrolling, clicking and scraping articles categorized by topic. The scraping function extracted the title, category, and legitimacy of each article.

Dataset Compilation

After scraping the data, it was compiled into CSV files, with each file containing articles belonging to a specific category. These CSV files were then combined into a single dataset using Pandas, a Python data manipulation library.

Dataset Statistics

Total Number of Articles

The combined dataset contains a total of 37,784 articles.

Distribution Across Topics

The articles are distributed across eight main topics, with the following frequency:

- أخبار (News): 3,156 articles
- ترفيه (Entertainment): 501 articles
- تكنولوجي ا (Technology): 405 articles
- ثقافة وف ن (Culture and Arts): 5,841 articles
- رياض ة (Sports): 16,984 articles
- سياس ة (Politics): 7,410 articles
- صحة (Health): 3,458 articles
- موسي ق (Music): 29 articles

Legitimacy Distribution

The majority of the articles are labeled as fake (زائ ف), with a frequency of 36,986 articles. Only 798 articles are labeled as true or legitimate.

Conclusion

In conclusion, this report outlines the methodologies used to create an Arabic dataset for fake news detection. The dataset consists of 37,784 articles across various topics, with a focus on identifying fake news. Such datasets are essential for training machine learning models to automatically detect and combat misinformation in Arabic-language content.