Summary Report: Web Scraping News Articles

Objective:

The objective of this project was to create a dataset for a future fake news detection project. The dataset will be used to train an AI model to distinguish between fake and real news, enabling users to authenticate articles or utilize the data for various data analysis projects.

Introduction:

To accomplish this task, we utilized several Python libraries including Beautiful Soup and Requests for web scraping, Pandas for data handling and cleaning, as well as the "re" library for pattern recognition. We identified multiple websites for collecting both real and fake articles:

Real Articles:

BBC Arabic Website: (https://www.bbc.com/arabic/)

Youm7 Website: (https://www.youm7.com)

Fake Articles:

Fatabyyano Website: (https://fatabyyano.net)

Verify-Sy Website: (https://verify-sy.com)

For each website, we systematically scraped information from all available articles. Please refer to the statistics below for detailed information about the gathered data.

Fake News Collection:

In our approach to gathering fake news articles, we observed that titles often contained phrases questioning the authenticity of the articles or explicitly labeled them as fake. Consequently, we adopted a different approach for scraping fake news. We extracted the links to the articles and accessed the webpages directly to scrape the original text of the fake news articles.

Fatabyyano Website:

We iterated through each page in the "زائف" category, extracted the links of each article card, accessed the links, and scraped data from the divs with the class 'ob-post-text'. The scraped data was then saved into a CSV file.

Verify-Sy Website:

Similarly, we iterated through each page, extracting article card descriptions with the class 'content-list-desc' and using them as the original fake news texts.

Real News Collection:

Real news articles were rich on the selected websites, with many articles featuring summaries in their titles, facilitating data collection.

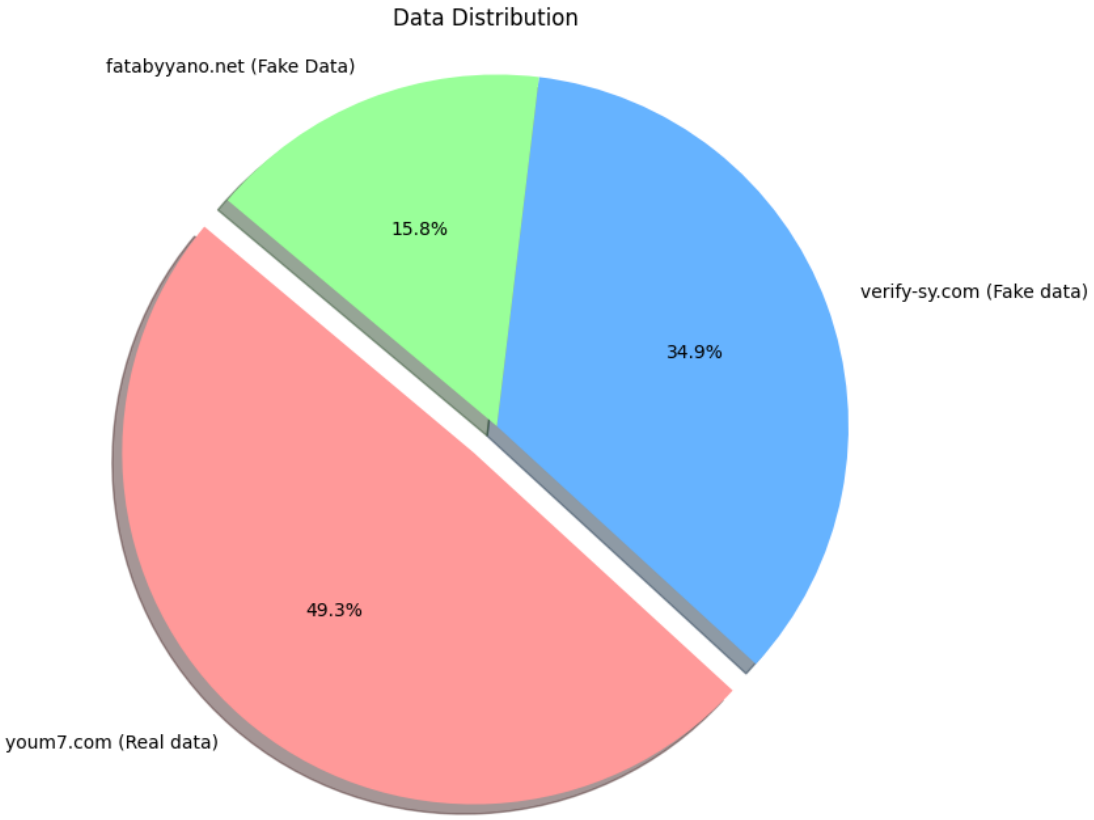
BBC Arabic Website:

We navigated through all pages in the "أخبار" section, extracted the HTML code, and retained text within divs with the class name "promo-text". We subsequently removed tags and their contents containing the words "المدة" for cleaning purposes.

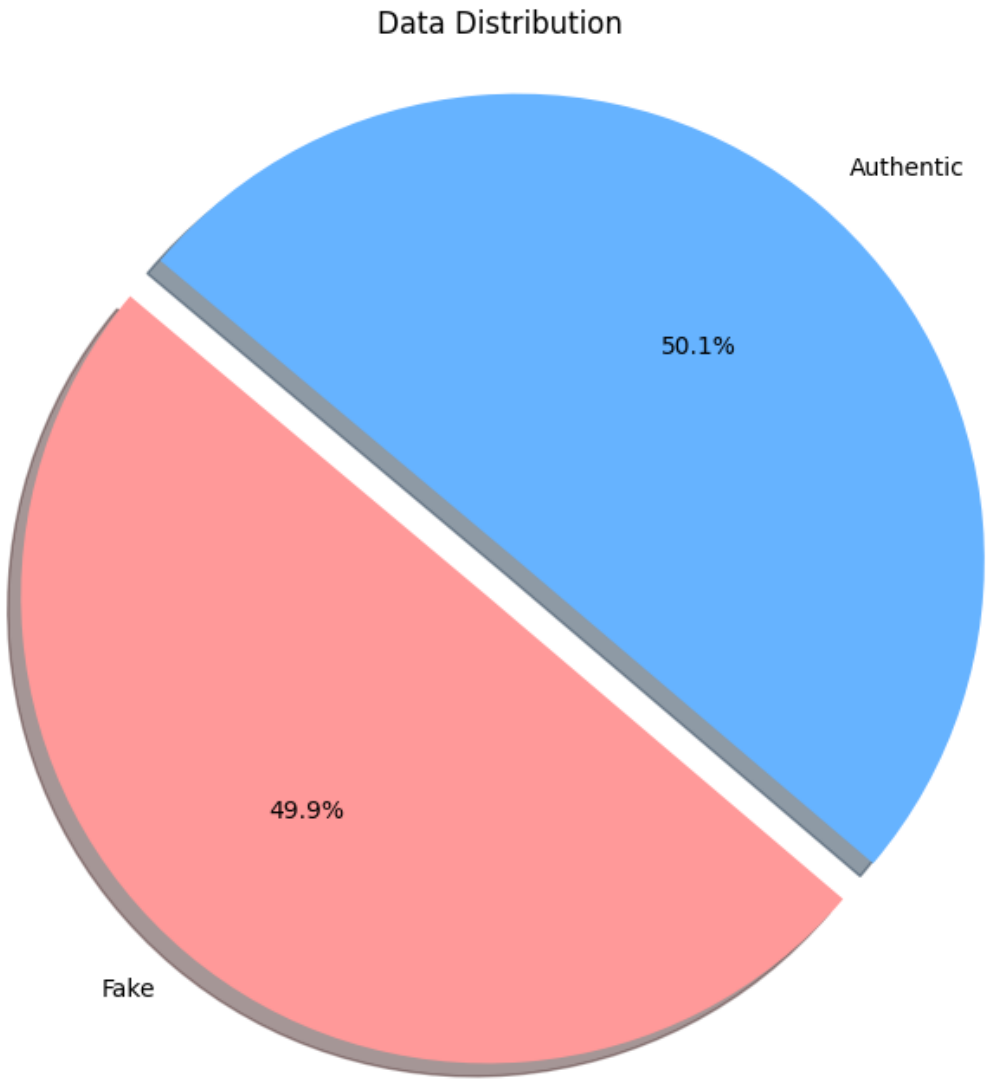
Youm 7Website:

This website provided the most extensive data, comprising thousands of pages, each containing numerous articles. We selected a representative sample of articles from various categories such as "تحقيقات" ,"اقتصاد" ,"حوادث" ,"تقارير" ,"سياسة," and "رياضة". After scraping all article titles, we compiled them into a single array and saved them as a CSV file.

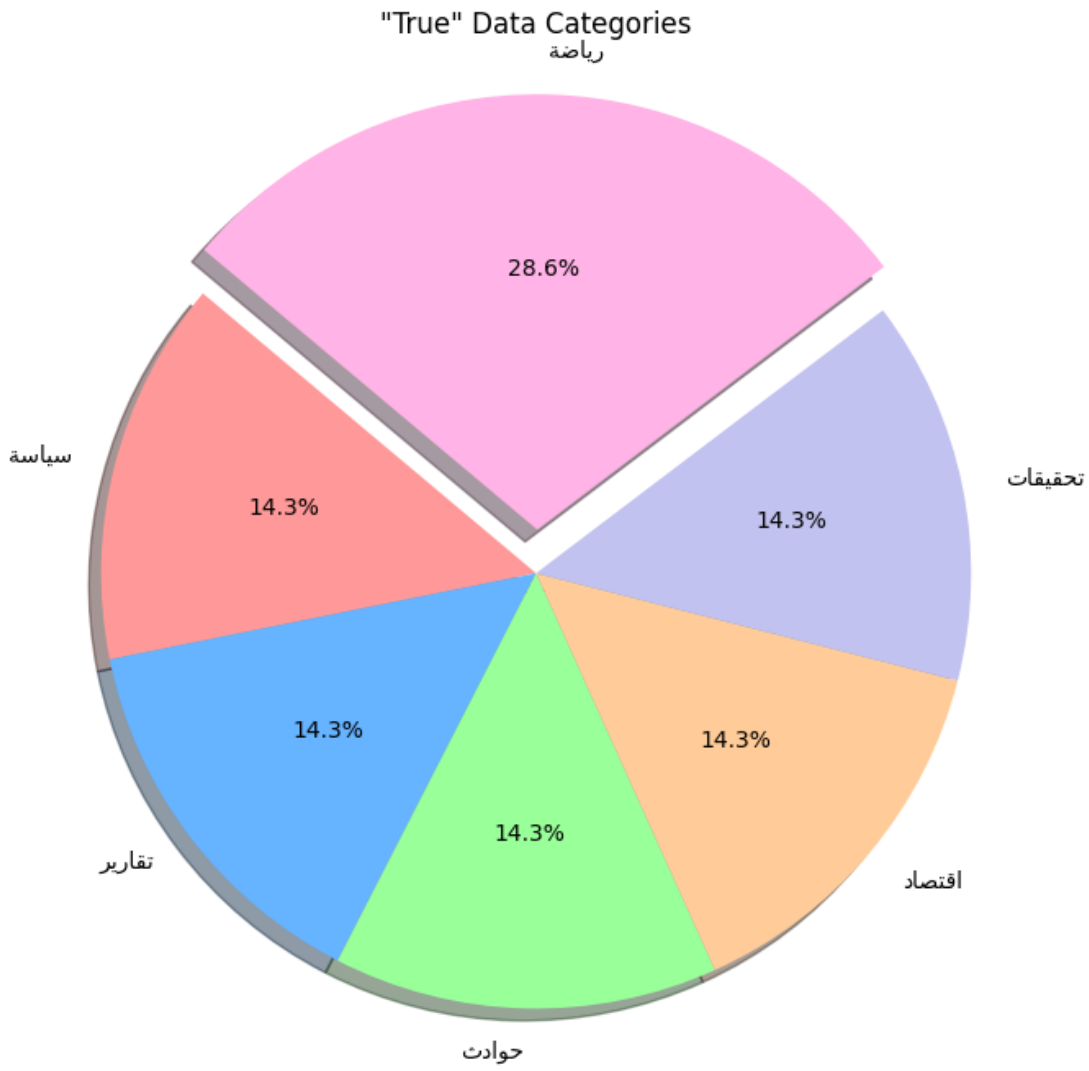
Statistics



Fake News Sources Distribution: Fatabyyano Website (15.8%), Verify-Sy Website (34.9%), Youm 7Website (49.3%)



Authenticity Distribution: 50.1% authentic, 49.9% fake news



Category Distribution: All categories ("سياسة," "تقارير," "حوادث," "اقتصاد," "تحقيقات") have a distribution of %14.3, except for "رياضة", which has a distribution of %28.4.

Conclusion:

In summary, our project achieved its goal of creating a valuable dataset for spotting fake news and analyzing news trends. By scraping articles from different websites using Python tools, we gathered a good mix of real and fake news samples. This dataset is diverse and comprehensive, offering plenty of insights for improving AI models and understanding media dynamics. It's a resource that could help us all become savvier news consumers in today's digital world.