الكلية متعددة التخصصات تارودانت

FACULTE POLYDISCIPLINAIRE TAROUDANT

Process Book

Project: Moroccan Media Landscape

|Module : Big Data 2
|Réaliser par : Aya KHIRANIA - Nidal Salima AZZAM
|Encadrer par : Pr. HAJJI et M. AIT BAHA
|Promotion : Master BDIA S3 2023-2024

Contents

Overview

This Media Analysis project delves into assessing media bias and objectivity through an extensive dataset comprising over 1 million Arabic news articles from prominent Moroccan media outlets.

Shifting our analytical focus towards the narratives and framing embedded within these articles, the project aims to uncover the subtleties of how news is presented and interpreted in the Moroccan media landscape.

In essence, while the project's direction shifted from its original sentiment analysis focus due to dataset limitations, this challenge opened the door to a broader exploration of media analysis.

Through inventive NLP techniques and the strategic use of Power BI for data visualization, we succeeded in achieving our revised objective: to conduct a deep dive into the narratives and representations within Moroccan electronic news media, offering valuable insights into their dynamics without the need for labeled data.

Scope of the Project

Data Sources and Types

The project relies on a dataset of over 1 million Arabic news articles sourced from 11 prominent electronic news sources named : **MNAD/v2**[1]**.**

The dataset is made available to the academic community for research purposes, such as data mining (clustering, classification, etc.), information retrieval (ranking, search, etc.), and other non-commercial activities.

Dataset Fields

- **Title:** The title of the article
- **Body:** The body of the article
- **Category:** The category of the article

[1] Multimodal News Article Dataset

- **Source:** The Electronic Newspaper source of the article

Interest and Limitations:

The primary sources include :

- **Hespress.ma:** As a major news outlet, Hespress contributes significantly to public discourse. However, its limitation lies in potential bias due to its popularity.
- **Alyaoum24.com:** Known for its diverse coverage, Alyaoum24.com adds richness to the dataset, yet its potential limitation includes a focus on sensational news.
- **Le360.com:** With a focus on in-depth analysis, Le360.com provides valuable insights, but its limitation may be a lower volume of articles.
- **Barlamane.com:** Specializing in political news, Barlamane.com offers a unique perspective, but its limitation may be a narrow thematic focus.
- **Alayam24.com:** Known for timely reporting, Alayam24.com enhances the dataset, but potential limitations include brevity in articles.
- **Al3omk.com:** Providing a regional perspective, Al3omk.com enriches the dataset, but its limitation may be a regional bias.
- **Anfaspress.com:** With a focus on varied topics, Anfaspress.com adds diversity, but its limitation may be less mainstream visibility.
- **Hibapress.com:** Contributing to regional news, Hibapress.com enriches the dataset, but its limitation may be a regional thematic focus.
- **Akhbarona.ma:** Diversifying the dataset with societal news, Akhbarona.ma's limitation may be a potential overlap with Hespress.
- **Medi1News.com:** Offering an international perspective, Medi1News.com adds breadth, but its limitation may be a lower volume of articles.
- **SnrtNews.com:** Contributing to national news, SnrtNews.com enriches the dataset, but its limitation may be a potential overlap with Hespress.

Tools and Technologies Used

Programming language :

Python:

Python's role in this project underscores its capacity to handle complex data analysis tasks and adapt to the specific needs of a project, proving it to be an indispensable resource in the exploration of media bias and objectivity.

Libraries and frameworks :

At the outset of our Media Analysis project, we initially employed a variety of libraries, including Polars for high-performance data manipulation, Vaex for efficient data handling, Farasa and Qalsadi for Arabic language processing, and bert-base-arabertv02 for deep learning-based text analysis.

Vaex:

The key features of Vaex, such as lazy evaluation, out-of-core processing, and efficient expressions, make it suitable for working with large datasets that might not fit into memory.

Farasa: used as a segmenter and Stemmer

FarasaSegmenter is an essential tool for Arabic text processing, specifically renowned for its robust tokenization capabilities. As a part of the Farasa suite developed by the Arabic Language Technologies (ALT) research group at the Qatar Computing Research Institute (QCRI), FarasaSegmenter excels in accurately breaking down Arabic text into meaningful units, such as words and subwords

FarasaStemmer, also part of the Farasa suite, plays a crucial role in Arabic text processing by focusing on the task of stemming. Stemming involves reducing words to their base or root form, aiding in tasks like information retrieval and text mining.

Qalsadi: Lemmatizer is a powerful tool designed for Arabic text processing, specializing in lemmatization—a fundamental task in natural language processing. Developed by the Arabic Language Technologies (ALT) research group at the Qatar Computing Research Institute (QCRI).

bert-base-arabertv02: short for Arabic Bidirectional Encoder Representations from Transformers, is a state-of-the-art language model developed by the Applied Artificial Intelligence Research Group at the University of Bahrain.

Polars: is a high-performance DataFrame library, designed to provide fast and efficient data processing capabilities. Inspired by the reigning pandas library, Polars takes things to another level, offering a seamless experience for working with large datasets that might not fit into memory.[2]

However, we encountered several challenges, particularly in presenting the results following the preprocessing phase. These issues prompted a strategic pivot in our approach.

We transitioned to using NLTK, a comprehensive library for natural language processing that better suited the linguistic nuances of our dataset, and tqdm for efficiently tracking the progress of our data processing tasks.

This shift significantly improved our workflow, enabling us to conduct more effective and insightful analysis of media bias and objectivity within the vast corpus of Moroccan news articles.[3]

NLTK :

NLTK is a standard python library with prebuilt functions and utilities for the ease of use and implementation. It is one of the most used libraries for natural language processing and computational linguistics.

tqdm :

Tqdm is a popular Python library that provides a simple and convenient way to add progress bars to loops and iterable objects. It gets its name from the Arabic name taqaddum, which means 'progress.'[4]

Project Goals and Objectives

The initial ambition of our Media Analysis project was to conduct a thorough investigation into sentiment analysis and media bias & objectivity across an extensive dataset of over 1 million Arabic news articles from various prominent Moroccan

[2] https://realpython.com/polars-python/#the-python-polars-library

[3] https://www.mygreatlearning.com/blog/nltk-tutorial-with-python/

[4] https://www.analyticsvidhya.com/blog/2021/05/how-to-use-progress-bars-in-python/#:~:text=Tqdm%20is%20a%20popular%20Python,taqaddum%2C%20which%20means%20'progress.

electronic media outlets. This goal was inspired by the successes of prior initiatives in sentiment analysis, aiming to broaden the analysis to a wider array of news sources.

The primary objective revolved around analyzing the textual data to unearth insights regarding the sentiments expressed in the articles, assessing the content's objectivity or subjectivity, and discerning the sentiment polarity—be it positive, negative, or neutral.

However, we encountered a significant challenge early in the project: the chosen dataset did not include labeled articles that could directly facilitate sentiment analysis or a straightforward assessment of media bias and objectivity, which were our principal targets. This limitation necessitated a strategic redirection of our efforts.

We decided to pivot towards a more nuanced media analysis using Natural Language Processing (NLP) techniques to extract as much information as possible from the articles, despite the absence of dates or any statistical values that could have enriched our analysis.

To adapt to these constraints, we leveraged NLP processes to mine the dataset for valuable insights, focusing on the narrative and framing within the articles rather than explicit sentiment labels. This approach allowed us to explore the underlying themes, tones, and potential biases present in the content without the need for predefined labels.

The culmination of our analysis was the extraction of a CSV file that was meticulously prepared to be compatible with Power BI. This strategic move enabled us to present our findings through an interactive dashboard, offering a comprehensive and engaging visualization of our analysis.

The dashboard serves as a powerful tool to convey the nuanced conclusions we drew about the portrayal of news and public discourse by Moroccan electronic newspapers.

Methodology

Data Loading and Exploration:

The project began by acquiring a large-scale Arabic news articles dataset from Hugging Face, consisting of over 1 million articles sourced from prominent electronic news outlets.

This dataset, hosted on the Hugging Face platform, served as a valuable resource for our sentiment analysis endeavor. Since the dataset was not pre-labeled, the initial phase involved loading it into VSCode Notebook for thorough exploration.

We delved into the diverse topics and sources represented in the dataset, gaining insights into the breadth of news coverage and the spectrum of electronic newspapers contributing to the corpus.

Text Preprocessing:

A critical step in preparing the dataset for the analysis was text preprocessing. Leveraging the power of NLTK, we addressed challenges related to the sheer volume of articles. By the end, we decided to use 100,000 articles instead of 1 million.

The preprocessing pipeline included tokenization, removal of affixes and punctuation, elimination of Arabic stopwords, and the application of stemming and lemmatization techniques. These steps ensured that the text data was standardized and conducive to accurate sentiment analysis.

This iterative process, conducted within the VSCode Notebook environment, underscored the significance of text preprocessing in managing the vast dataset and paved the way for subsequent phases of subjectivity/objectivity classification.

Outcome

To analyze media bias and objectivity in Moroccan electronic newspapers, it's essential to design graphs that allow for clear visualization and deep understanding of trends and patterns in article writing.

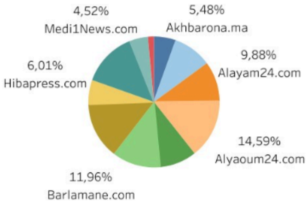
Her is the final created Dashboard:

Moroccan Media Landscape

Total sources

11

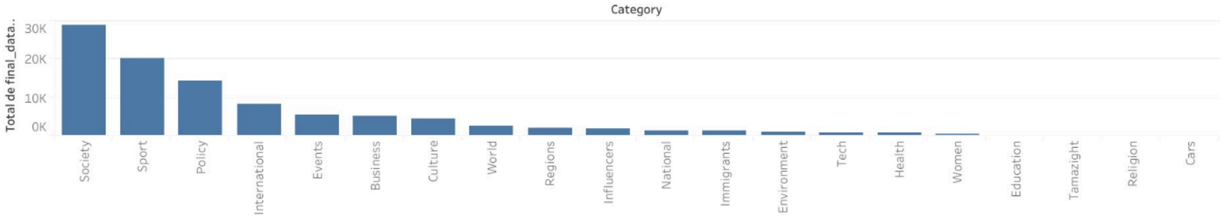
Source wise number of news



Total de final_data.csv

100 000

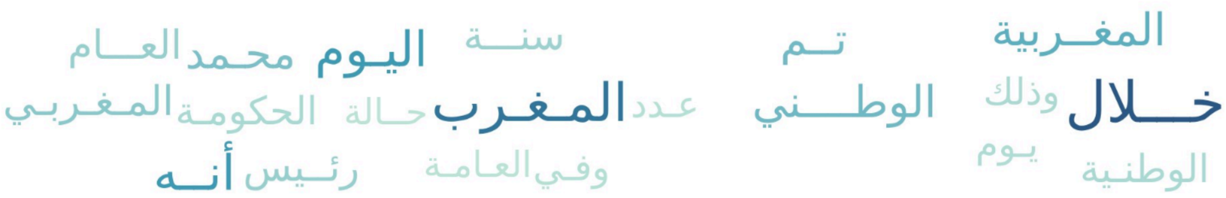
Category wise number of news



Total news

100 000

Top 20 words used



Appear Time

24 056 62 485

Category and Source wise news count

Source	Business	Cars	Culture	Education	Environment	Events	Health	Immigrants	Influencers	International	National	Policy	Regions	Religion	Society	Sport	Tamazight	Tech	Women	World
Akhbarona.ma	296	2	196	130		1818	328			818	445	655		49		375		285	83	
Al3omk.com	602		463									2245			4894	1231				
Alayam24.com	563		147			554			382	1135		2687			2604	1806				
Alyaoum24.com	287	3	630		301	346	333	107	11	859		1798	1		6151	3741		20		
Anfaspress.com	844		743			440		250	313	921		1619			2923	1070				
Barlamane.com	479		732		438					2618		1732			4041	1922				
Hespress.ma	1066	33	696		401	1841		347				1264	2126		1910	1558	109	60		2546
Hibapress.com	271		242			601		83		1287		856		28	2106	539				
Le360.com	593		335					553	1037	501	916	1352			3179	4718			383	
Medi1News.com			320												1030	2868		305		
SnrtNews.com	180									236		191			291	539		37		

This dashboard presents a comprehensive analysis of the Moroccan media landscape. It visualizes data from 11 news sources, offering insights into news distribution across various media outlets and categories. The pie chart in the upper right shows the source-wise distribution of news, with Alyaoum24.com contributing the highest percentage of articles.

In the category-wise number of news bar chart, 'Sports' emerges as the most reported topic, followed by 'Society' and 'Politics', indicating a strong interest in these areas within the Moroccan media space. The bottom section of the dashboard highlights the frequency of the top 20 words used in the news coverage, reflecting the current hot topics and the prevalent discourse in Moroccan media. Words like "المغربية" (Moroccan), "سنة" (year), and "الوطني" (national) appear prominently, suggesting a focus on national events and temporal markers in reporting.

The 'Category and Source wise news count' matrix further breaks down the number of articles per category for each news source, revealing how different outlets prioritize news topics. For instance, Hespress.com has a broad coverage across categories, while other sources may have more focused content areas.

Conclusion

The extracted dashboard provides a valuable macro-level snapshot of media focus in Morocco, which can be utilized by analysts, media planners, and researchers to understand media trends, gauge public interest, and assess the diversity of news coverage across different domains.