*future internet*

MDPI

*Article*

# A Methodological Framework for AI-Driven Textual Data Analysis in Digital Media

**Douglas Cordeiro [1,*], Carlos Lopezosa [2] and Javier Guallar [2]**

[1] Faculty of Information and Communication, Federal University of Goiás, Goiânia 74690-900, GO, Brazil
[2] Faculty of Information and Audiovisual Media, University of Barcelona, 08193 Barcelona, Spain; lopezosa@ub.edu (C.L.), jguallar@ub.edu (J.G.)
* Correspondence: cordeiro@ufg.br

**Abstract:** The growing volume of textual data generated on digital media platforms presents significant challenges for the analysis and interpretation of information. This article proposes a methodological approach that combines artificial intelligence (AI) techniques and statistical methods to explore and analyze textual data from digital media. The framework, titled DAFIM (Data Analysis Framework for Information and Media), includes strategies for data collection through APIs and web scraping, textual data processing, and data enrichment using AI solutions, including named entity recognition (people, locations, objects, and brands) and the detection of clickbait in news. Sentiment analysis and text clustering techniques are integrated to support content analysis. The potential applications of this methodology include social networks, news aggregators, news portals, and newsletters, offering a robust framework for studying digital data and supporting informed decision-making. The proposed framework is validated through a case study involving data extracted from the Google News aggregation platform, focusing on the Israel–Lebanon conflict. This demonstrates the framework's capability to uncover narrative patterns, content trends, and clickbait detection while also highlighting its advantages and limitations.

**Keywords:** digital media; natural language processing (NLP); text analysis; sentiment analysis; artificial intelligence; statistics.



Academic Editors: Arjun Mukherjee and Leonardo Ranaldi

Received: 13 December 2024
Revised: 24 January 2025
Accepted: 28 January 2025
Published: 3 February 2025

**Citation:** Cordeiro, D.; Lopezosa, C.; Guallar, J. A Methodological Framework for AI-Driven Textual Data Analysis in Digital Media. *Future Internet* **2025**, *17*, 59. https://doi.org/10.3390/fi17020059



## 1. Introduction

Advancements in the field of information and communication technologies have transformed social and technological dynamics, creating a scenario in which the mass production of data is not only an essential component of digital platforms but also an intrinsic feature of contemporary society's daily activities. In this context, digital media play a fundamental role in amplifying and facilitating communication, offering new possibilities for information exchange. According to [1], digital media can be defined as a set of practices and systems characterized by programmatic production and the integration of dynamic semiotic modes, aimed at fostering interactions mediated by technology. This framework is closely tied to access to personal technological devices and the widespread adoption of the internet [2].

Digital media not only expand the reach and speed of interactions, but also introduce a new participatory dimension to the media experience, enabling the creation of participatory materialities and reflecting the user's ability not only to consume content but also to influence and alter the media experience in immersive and dynamic environments. This transcends traditional one-way communication, facilitating the co-creation of meanings and

experiences [1]. The dynamic and interactive nature of digital media suggests a substantial reconfiguration of relationships between users and platforms, pointing to behavioral and structural changes in how resources, experiences, and information are managed [3].

The massive volume characteristic of a big data environment and the complexity of data generated on digital platforms have posed significant challenges for researchers across multiple disciplines [4,5]. Social media, blogs, newsletters, news portals, and content aggregators not only produce data on unprecedented scales but also reflect and shape individual and collective behaviors. However, existing methodological approaches are typically either overly generic, failing to address the specific characteristics of data originating from digital media, or overly specific, focusing on isolated cases for individual digital media platforms. As a result, such solutions are often unable to comprehensively and contextually address the challenges posed by the fragmented and unstructured nature of these data. Data generated within the context of digital media, alongside specific experiments, enable an understanding of how social indicators such as prestige and popularity interact with individual content evaluation and other trends [2]. Moreover, digital media analysis allows researchers to identify, classify, and quantify phenomena and patterns associated with behaviors, consumption, and power dynamics manifested in social media environments [6], or even in more specialized settings, such as digital journalism platforms [7,8].

On one hand, the distinctive structures of digital platforms provide a wealth of quantitative inputs that support the exploration of descriptive and exploratory studies, leveraging, for example, statistical analyses and data visualization [9–11]. On the other hand, interactions enabled by digital environments have fostered a surge in text-based communication [2], whether through the dissemination of originally informational texts or manifestations materialized as comments on digital platforms, making them valuable raw material for in-depth analysis. To address these challenges, our methodological approach proposes the use of specific techniques to bridge gaps in existing frameworks, emphasizing comprehensive data processing and analytical rigor. Textual data facilitate the understanding of social dynamics, information flows, and decision-making processes. However, the unstructured, fragmented, and contextual nature of these data demands the use of robust and complex strategies to extract indicators and identify patterns, often relying on solutions grounded in areas such as natural language processing (NLP).

Beyond serving as communication channels, digital platforms operate as ecosystems for the production and exchange of information, often modulated by algorithms [12]. This context creates conditions where not only the content itself but also the manner of its distribution influences processes of perception, consumption, and engagement. The systematic study of these platforms requires the adoption of methodological approaches that integrate data collection and processing technologies with robust analyses, drawing on solutions such as artificial intelligence (AI) and statistical analysis.

This article makes two primary contributions. First, it highlights how the proposed methodological approach enhances the analysis of textual data from digital media by presenting a framework focused on data enrichment and knowledge discovery. Second, it demonstrates the application of the proposed approach through a case study that showcases its ability to uncover insights into textual patterns and dynamics on digital platforms. The importance of a computationally robust methodological approach extends beyond the individual analysis of digital platform data. Considering large-scale applications, strategies that facilitate the analysis of massive datasets for collection and examination contribute to advancing knowledge about the underlying mechanisms of online behavior. In studies integrating social media data, for instance, the analytical procedures associated with our methodological proposal enable the understanding of processes such as polarization, misinformation, and public participation across various contexts. Another example of

application lies in the analysis of news aggregators, where the methodological approach supports the establishment of procedures in a highly dynamic digital environment. This environment demands the continuous application of data extraction routines and enables the consolidation of segmented and targeted indicators, aiding the comprehension of news phenomena and nuances in information distribution [13,14].

This article proposes a methodological approach that integrates AI techniques within the context of natural language processing and statistical methods for the exploration and analysis of textual data from digital platforms. The proposal includes strategies for data collection (via APIs and web scraping), preprocessing, enrichment, and analysis, with a particular focus on named entity recognition, sentiment analysis, and text clustering. The results underscore the advantages of the proposed approach over specific limitations of other methodological alternatives, such as its ability to integrate datasets and its potential to enhance textual data enrichment. The practical relevance of the approach is demonstrated through a case study that illustrates its advantages and limitations, providing insights into the potential and challenges of an integrated analysis based on large-scale data.

## 2. Background

The analysis of digital media data, particularly in contexts involving the processing of large volumes of textual data, requires robust and well-defined methodological strategies that support the entire workflow, from initial data collection to analysis and interpretation, ensuring the extraction of information relevant to the research objectives. While classical approaches such as KDD (Knowledge Discovery in Databases) [15] and CRISP-DM (Cross-Industry Standard Process for Data Mining) [16] remain important frameworks for knowledge discovery processes [17–19], their application within the context of digital media demands alignment with the specific characteristics of these platforms, such as the predominant use of natural language and the diversity of data formats.

The use of data mining and artificial intelligence algorithms for conducting quantitative studies in digital media has grown exponentially. This increase is primarily due to the indispensable role of these techniques in tasks such as detecting fake news and evaluating news credibility [20–23], studying the personalization of informational content [24], applying sentiment analysis methods, and detecting sensationalist news [25,26]. Additionally, they are crucial for monitoring news, both for specific datasets [27,28] and for specific topics [29,30].

Regarding the detection of fake news, the specific use of algorithms and AI has become widespread as a technique for classifying deceptive content, particularly on social media and digital media platforms. In these cases, various strategies have been developed, including supervised algorithm-based techniques capable of achieving high accuracy in data collection and interpretation through decision trees [20]; machine learning and deep learning techniques combined, for instance, with the Apache Spark framework [21]; and hybrid approaches that integrate technologies such as support vector machines and natural language processing [23]. Additionally, studies comparing systems based on artificial neural networks and the BERT algorithm for evaluating news source credibility have confirmed the utility of advanced semantic models as an effective method for studying fake news [22].

With respect to the personalization of content through data mining and artificial intelligence algorithms, some studies have explored how wireless detection technology and enhanced algorithms predict user acceptance of personalized news sites [24], confirming that these models are capable of tailoring news to user preferences.

Another prominent research focus in this area has centered on the use of technology as a tool for conducting sentiment analysis and detecting sensationalist content. To this end,

some researchers have developed systems capable of identifying sensationalist headlines through natural language processing and feature engineering, which involves the extraction and transformation of variables from raw data [25]. Furthermore, neural network models have been implemented for named entity recognition and emotional viewpoint analysis in online news [26]. These AI-based data collection methods have, on the one hand, enabled the classification of emotions with improved accuracy compared to traditional methods [30]. On the other hand, they have also challenged the perception of certain negative biases that may arise from manual data collection processes [29].

Another key area in quantitative studies on digital media that has been significantly impacted by the use of AI is related to text analysis and data mining. In this context, some studies have employed topic models and specific algorithms, such as the Apriori algorithm, to extract knowledge from news content and improve media monitoring reports [27]. Additionally, news mining systems have been developed based on both the Internet of Things and sensor fusion, achieving the extraction of specific data with an accuracy of 98% [28].

One of the fundamental steps in achieving robust results involves obtaining data that are sufficiently representative for analyzing the problem under investigation. The massive and growing volume of data generated on digital media platforms such as social networks, news portals, and content aggregators offers a vast field of study for research focused on exploring social behaviors, analyzing communication patterns, and identifying cultural, political, social, and economic trends. These data, which include text, images, videos, and user interactions (such as likes, shares, and comments), constitute rich sources of information on opinions, interests, and dynamics within virtual communities. Conducting analytical procedures on these datasets can enable the generation of deep and valuable indicators for various fields. Ref. [31] points out that data collection represents both an enabler and an obstacle to the development of scientific research, reflecting challenges inherent to platform formats and the tools and techniques available for data collection. Unlike traditional approaches based on data collection through interviews, structured observations, focus groups, and surveys, digital media data require robust and specialized techniques [32] that allow for large-scale exploration, even at a global level, facilitating comparative research across different regions. The collection of digital media data is predominantly carried out through two strategies: APIs (Application Programming Interfaces) and web scraping techniques.

APIs, when available, provide a structured solution for accessing data but often limit the scope of information due to restrictions imposed by platforms [33]. One of the main advantages of using APIs is the ability to perform specific and customized queries within the databases of digital platforms, including access to data from posts, profiles, groups, interactions, and engagement metrics. Furthermore, data access via APIs is typically faster and more efficient than other approaches, as platforms develop their APIs for large-scale use, offering access to historical data and aggregated information that is not directly available through public interfaces, such as metadata from posts. Various digital media platforms, especially in the context of social networks, offer general-purpose APIs for data extraction, such as Bluesky's API, X's API, and Meta's Graph API. Additionally, there are third-party-maintained APIs, such as the Google News API, managed by SerpAPI.

Despite the potential advantages offered by APIs, ref. [34] describes many APIs of digital media platforms as highly restrictive, a situation referred to by [35] as the "APIcalypse". In this scenario, data sharing among researchers and the use of web scraping solutions have become alternatives for conducting research related to digital platforms. On the other hand, legal actions such as the enforcement of the Digital Services Act (DSA) Article 40 by the European Commission (2023) aim to provide solutions for enabling researchers to

request access to data from platforms classified as "Very Large Online Platforms" (VLOPs) and "Very Large Online Search Engines" (VLOSEs). This initiative has driven updates to transparency policies, including the availability of APIs specifically designed for research purposes, such as TikTok's Research API and Meta's FORT Researcher API. Similarly, governments in other regions, such as the United Kingdom with the proposed HL Bill 40—Data (Use and Access) Bill, have sought comparable solutions to support research development based on large datasets.

Regions like Latin America lack specific legislation regarding the provision of digital platform data for scientific research. Furthermore, while the reliability of data returned by APIs is technically an advantage, studies such as [36] highlight issues related to data consistency. A systematic audit of TikTok's Research API revealed discrepancies between the data provided by the API and the information directly available from posts accessed through the social network's application. These factors, combined with the restrictions imposed by major APIs, can compromise certain research endeavors, necessitating the use of alternative strategies, such as web scraping, for data collection.

The use of data extraction routines based on web scraping offers greater flexibility by directly accessing data from web pages, overcoming the barriers imposed by APIs or in scenarios where APIs are unavailable [37]. Web scraping is a data collection technique that involves the direct extraction of information from web pages [38]. Unlike APIs, which provide access to structured and often aggregated data, web scraping collects data as they are configured and presented on the pages, enabling the capture of more detailed and specific information that, in many cases, is not accessible through APIs. This technique is particularly useful when a platform does not offer an accessible API or when the data of interest are not readily available.

In the context of digital media, web scraping is used to capture data from social networks [39–41], news portals [42], blogs [43], and other dynamic sites [44]. This technique enables the collection of textual data, such as post content, articles, and comments, as well as metadata such as dates, publication times, interaction counts, and geotags associated with the content. Web scraping provides significant flexibility regarding the data collected, allowing researchers to obtain information in formats that APIs do not offer or that are limited by commercial restrictions.

Despite its numerous advantages, the use of web scraping for data collection on digital platforms raises ethical, legal, and privacy concerns that must be carefully evaluated to ensure responsible practices aligned with applicable regulations. These concerns include legal ambiguities stemming from varying legislation across jurisdictions and the absence of a common ethical framework to guide the use of data extracted from digital platforms [45]. This complex and sensitive landscape demands particular attention to both legal compliance and ethical responsibility.

Data extraction through web scraping must respect individuals' privacy rights, especially when personally identifiable information (PII) is accessible [46,47]. Regulations such as the General Data Protection Regulation (GDPR) in the European Union and Brazil's General Data Protection Law (Lei Geral de Proteção de Dados, LGPD) set strict guidelines regarding the use and protection of personal data. Under these regulations, personal data may only be processed with an appropriate legal basis, such as explicit consent, legitimate interest, or compliance with legal obligations. Furthermore, web scraping can place excessive load on the servers of digital platforms, potentially impairing their functionality and performance, which may result in technical issues and constitute a misuse of platform resources.

Building on this, as explored by [48], it is essential to consider certain best practices when implementing a data collection approach based on web scraping: minimizing tech-

nical impact by configuring scripts to operate with appropriate time intervals, thereby avoiding server overload; ensuring compliance with local laws and regulations as well as platform terms of use; anonymizing collected data and limiting the use of personally identifiable information; and assessing the benefits and potential drawbacks of data collection, prioritizing integrity and social responsibility in research.

Once the data are obtained, specific treatment and processing procedures must be applied to enable the generation of indicators that support the analyses of interest. In our proposal, we suggest the use of natural language processing techniques, such as named entity recognition (NER), and pre-trained Large Language Models (LLMs), such as BERT, GPT, and LLaMa. Although the proposed framework allows for the use of different LLMs, we chose to work with BERT in our experiments because it offers pre-trained versions in various languages and domains, provides high flexibility for adaptation to specific tasks, and strikes a balance between performance and computational cost.

NER is a natural language processing (NLP) technique that identifies and classifies specific elements within a text [49]. This process facilitates the extraction of structured information from unstructured data, generating datasets that are semantically organized and enriched. In the context of digital media, NER can be utilized for tasks such as analyzing communication strategies [50,51] and monitoring phenomena and events [14]. Applying NER to large volumes of textual data, especially in digital environments characterized by informal, fragmented, ambiguous, or highly specific language, requires robust models to ensure precision and relevance. BERT is a language model whose primary feature is its ability to process a word's context by considering both its preceding and succeeding terms—known as bidirectional learning [52]. BERT is widely used in NLP tasks, including those utilizing NER-based models. By integrating NER with BERT, it is possible to take advantage of the contextualization present in data from digital platforms, significantly improving accuracy in identifying entities within complex texts [53,54]. It is worth noting that other Large Language Models can be utilized; however, in our proposal, we opted to use models from the BERT family due to the favorable results observed, as highlighted in Section 4.

## 3. The Proposed Approach: Data Analysis Framework for Information and Media

Figure 1 illustrates the overall architecture of our methodological proposal, titled DAFIM (Data Analysis Framework for Information and Media), which organizes the process into three main stages: data extraction, data enrichment, and knowledge discovery. In the first stage, data extraction, the focus is on extracting information directly from web sources. Next, in the data preprocessing stage, the collected data undergo processes such as cleaning, normalization, transformation, and the identification of derived attributes. Finally, the knowledge discovery stage involves the application of data mining and artificial intelligence algorithms for more sophisticated analyses, primarily in the areas of text classification and clustering. The following subsections provide details on each step of the framework.

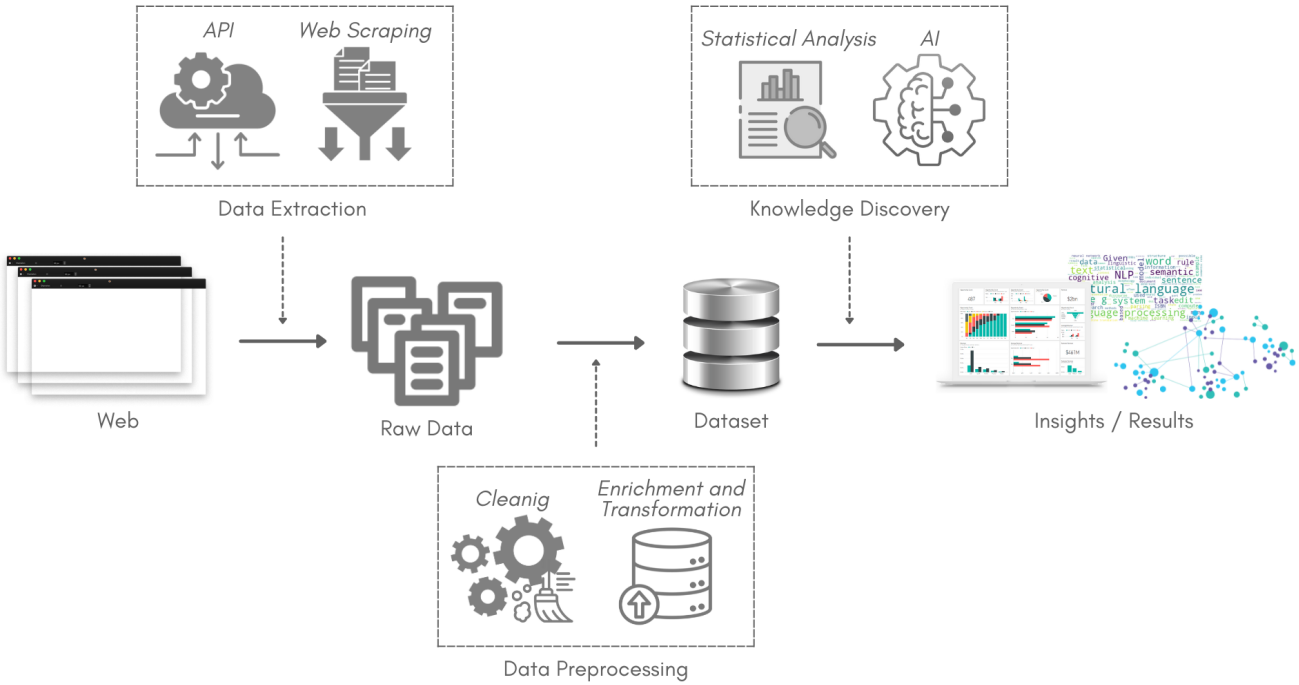



**Figure 1.** General architecture.

### *3.1. Data Extraction*

Figure 2 presents the general framework for data collection using web scraping routines. We used BPMN notation in constructing the diagrams to facilitate understanding of the procedures performed [55]. Data collection on digital platforms often involves navigating pagination structures, a necessary step to access the addresses of posts organized within higher hierarchical pages. These posts, which may include journalistic texts, blog entries, or social media content, are typically organized in a structured manner and are frequently presented in paginated formats that are interactively navigable. Pagination may be implemented through navigation buttons, page counters, or infinite scrolling, each of which dictates the specific strategies required for effective data collection.

On digital platforms where direct online data collection is possible, a web scraping routine can be programmed to access pages and simultaneously execute pagination and extract the desired data, such as the URLs leading to pages containing the final data of interest. Conversely, in situations where direct URL collection from pagination structures is not feasible, alternative approaches must be adopted. One solution involves downloading the pages of interest using commands executed in scripting environments, such as Bash, facilitated by browsers that simulate human interactions. From there, the extraction of URLs can be performed on offline files. In both cases, the set of URLs can be stored in a semi-structured file format, such as CSVs (comma-separated values), simplifying subsequent manipulation and analysis. Regardless of whether the collection is performed online or offline, both approaches converge in a common stage of organizing and structuring the target URLs.

Data extraction from pre-identified URLs requires an initial analysis to verify the feasibility of direct collection through web scraping. Some digital platforms have structures that complicate or prevent direct extraction, necessitating alternative procedures such as downloading the pages of interest. This process is similar to that described in the context of data collection from pagination structures, where pages are downloaded for offline access. Whether access is conducted online or offline (based on previously downloaded pages), the effectiveness of implementing a scraping solution depends on a detailed structural analysis

of the page. This analysis aims to identify markers and structural elements (such as CSS classes, HTML attributes, or patterns in the DOM) that enable the configuration of a script capable of locating and extracting the data of interest.

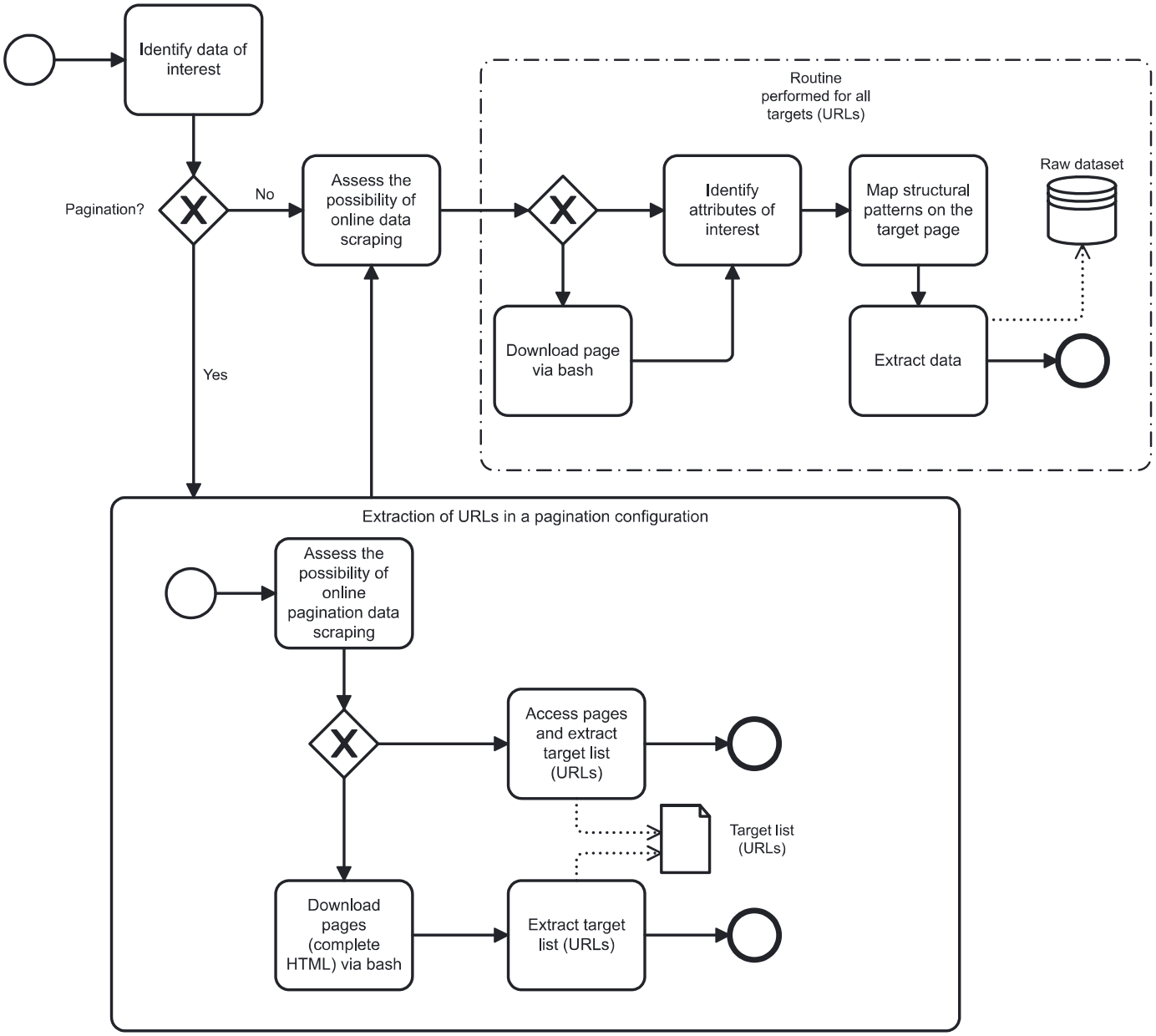



**Figure 2.** Web scraping data extraction scheme. Note: the diamond symbol with an "X" inside represents an exclusive OR (XOR) logical operation.

Through the development of a data scraping script, the specific information from each page is automatically collected, and the raw data obtained are organized and stored in an integrated database, ready for the subsequent stages of processing, enrichment, and analysis. This workflow ensures systematicity and consistency in the data collection process, even when faced with technical limitations imposed by certain digital platforms. The choice between direct data scraping strategies or page download methods depends on the constraints set by the platforms, such as access restrictions or the complexity of data organization structures at the source. However, each strategy provides mechanisms to overcome technical barriers and scale data collection effectively.

### *3.2. Data Preprocessing and Enrichment*

After extracting the data of interest, it is necessary to conduct verification and preprocessing routines to ensure data quality and maximize its analytical value. Numerical checks and routines based on regular expressions are particularly useful for data cleaning and enrichment procedures, such as identifying occurrences of relevant terms, including links, social media mentions, or hashtags. Furthermore, the use of artificial intelligence techniques associated with named entity recognition models (such as persons, geographic locations, objects, and brands) enhances data enrichment by enabling a deeper understanding of the semantic context of textual content [56]. Figure 3 illustrates the proposed framework for applying data preprocessing and enrichment routines.

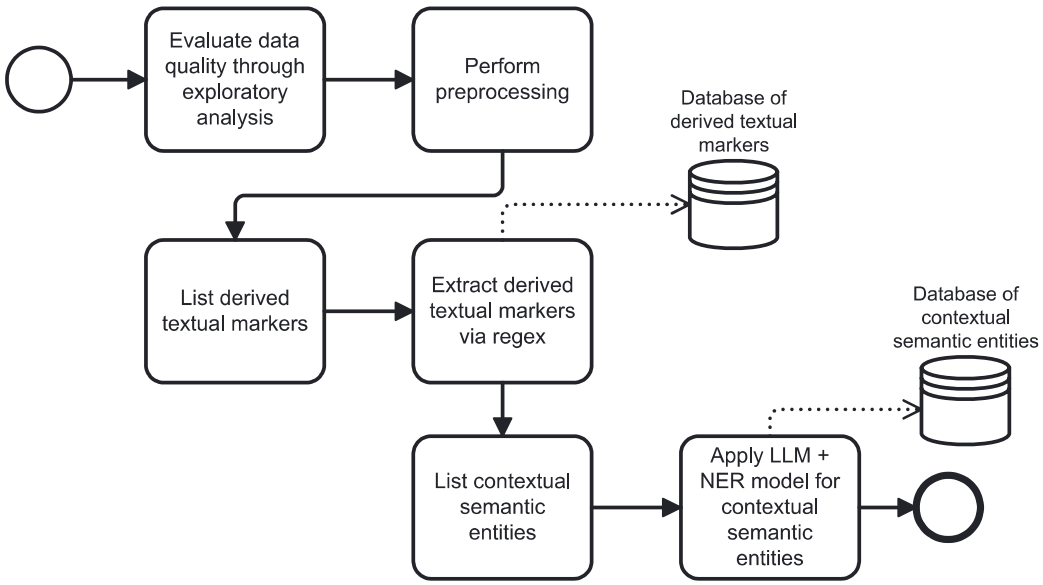



**Figure 3.** Data preprocessing and enrichment scheme.

Following this approach, Exploratory Data Analysis (EDA) can be employed to assess the quality of the collected data, enabling the investigation of the semantic and structural properties of the dataset [57]. One of the primary challenges in data analysis is mitigating biases stemming from datasets that are insufficiently representative or exhibit anomalies, such as inconsistent, missing, or extreme values (outliers). EDA, through techniques such as correlation matrices, histograms, boxplots, and frequency analysis, allows for the verification and, if necessary, the treatment of candidate data for analysis, ensuring sample representativeness.

An essential aspect, particularly when dealing with textual data, is the use of natural language processing (NLP) techniques to verify semantic consistency. These techniques facilitate, for example, the identification of discrepancies in the categorization of entities or in the standardization of terms, guiding the refinement of preprocessing procedures such as cleaning and normalization [58,59]. Within this context, actions are implemented to ensure the uniformity and quality of the corpus. One of the initial steps involves normalizing text capitalization to address inconsistencies arising from variations in the writing of specific terms. Subsequently, the removal of special characters, including punctuation marks, symbols, and numbers with no relevant semantic value for the analysis, is carried out. This contributes to noise reduction and enhances the structural clarity of the data. It is important to emphasize that these procedures should align with the type of analysis to be performed and the context of the digital platform being studied, as structural differences exist in the composition of textual corpora from platforms serving distinct purposes. For instance, textual data from social media, where writing tends to be more informal and

spontaneous, contrasts with data from news portals, which generally adhere to stricter grammatical standards and formal writing conventions.

In the preprocessing stage of textual data, lemmatization is particularly relevant in contexts with high variability in writing styles, especially on social media. This technique reduces words to their base or root forms, enabling them to be treated as a single entity. Lemmatization is crucial for term-based analyses as it reduces redundancy and enhances semantic coherence. Additionally, the use of semantic embeddings, such as Word2Vec [60–62], GloVe [63], or an LLM like BERT [64], transforms words into vector representations, capturing semantic meaning in various contexts. This facilitates the application of models to measure semantic similarity between words and phrases, aiding in the detection of out-of-context or miscategorized terms, thus improving the accuracy and consistency of textual analyses.

The next stage involves identifying attributes derived from structured textual markers, which represents a key aspect of data enrichment. These attributes correspond to textual occurrences that follow predefined patterns, such as mentions and hashtags on social media or hyperlinks in general. The definition of the list of derived attributes depends on the specific context of the problem and the digital platform being explored. One approach to extracting these attributes is the use of routines based on regular expressions [39]. The extracted data are consolidated into derived datasets, which should exhibit a one-to-many cardinality relationship with the main database.

Following the identification and extraction of attributes from structured textual markers, it is essential to define, in accordance with the research problem and the context of the digital platform, attributes related to contextual semantic entities. These refer to unstructured markers such as names of individuals, brands, political parties, and geographic locations. For this purpose, an NER-LLM solution is employed. It is important to ensure that the use of pre-trained models adheres to best practices for validation and contextualization, which involve the use of testing and validation metrics. Based on the application of the NER-LLM solution, a new derived file is generated for each contextual semantic entity of interest, following a structure similar to that of datasets containing structured textual markers. These derived files maintain a one-to-many cardinality relationship with the primary dataset.

### *3.3. Knowledge Discovery*

Knowledge discovery is an iterative process aimed at identifying non-obvious patterns, relationships, and trends within large datasets. This process employs data mining techniques, such as classification, clustering, and association algorithms, alongside statistical models and machine learning approaches. The use of specific techniques must align with the purposes of the research being conducted and the context of the data. In other words, there are certain types of problems where the application of clustering or classification solutions is not necessary. Data visualization—through graphs, tables, and other visual representations—plays a critical role in exploring and communicating findings. Figure 4 illustrates the proposed process flowchart, in which statistical analysis and predictive modeling are conducted in parallel and interactively. Subsequently, the results are consolidated using tabulation and data visualization techniques.

The application of clustering algorithms, such as DBSCAN [65–67] and Deep Embedded Clustering (DEC) [68,69], as well as text classification techniques like Random Forest [70,71] and LLMs [72–75], such as BERT and GPT, plays a central role in knowledge discovery for digital media analysis. These techniques enable the extraction of useful and relevant information from large volumes of textual data, uncovering latent patterns and grouping information to facilitate interpretations tailored to various research contexts.

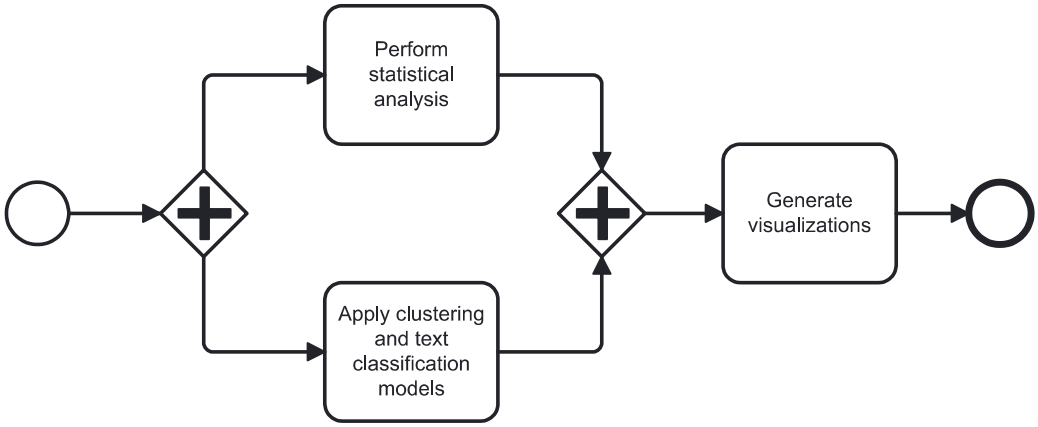



**Figure 4.** Knowledge discovery scheme. Note: the diamond symbol with a "+" inside represents the execution of both paths.

Following our proposed approach, models based on natural language processing (NLP) algorithms should be adapted to meet the specific needs of the research analyses, being selected and configured according to the identified research problem. For instance, sentiment analysis solutions, which measure public perception on specific topics [76–78], are particularly relevant for studies exploring public opinions. Conversely, methods focused on detecting fake news and clickbait [79–82] and topic analysis [83,84] are better suited for investigations aimed at identifying and understanding misleading content.

By overcoming the limitations of manual analysis, these approaches allow for the identification of emerging trends, audience segmentation, and the evaluation of communication strategies. However, it is crucial to acknowledge the limitations of these techniques, such as their challenges in capturing linguistic nuances, ambiguities, and the influence of contextual factors. These challenges can be mitigated through advancements in deep learning, particularly with the use of pre-trained language models, which offer enhanced contextual and semantic interpretation capabilities in textual analysis [85,86]. Additionally, the application of computational solutions must be guided by best practices in contextualization, testing, and validation of the models employed.

In conjunction, the integration of statistical techniques is employed to enhance the robustness of the analyses. Quantitative models enable the measurement of the significance of identified patterns, the validation of hypotheses, and the assurance that the inferences drawn are replicable. This combination of artificial intelligence and statistical methods facilitates the transformation of raw data into comprehensible indicators, providing analytical foundations both for the outcomes derived during the knowledge discovery stage and for the variables previously generated during the data enrichment phase.

The final stage of the process involves data visualization, which translates results into intuitive representations. In the context of NLP solutions, specific visualizations, such as similarity graphs for analyzing semantic relationships and word clouds for identifying the most frequent terms, highlight relevant textual patterns. Word clouds are visual representations that reflect the frequency of terms in a textual corpus, where terms are displayed in sizes proportional to their frequency. Similarity graphs, in addition to presenting the most frequent terms as nodes, also depict their relationships in terms of semantic association. This is represented by the co-occurrence of two terms within text segments, where the width of the edge indicates the strength of the association. Additionally, traditional numerical visualizations, such as bar charts, histograms, and scatter plots, can be employed when the study requires a quantitative approach. These representations support the communication of results and enable deeper insights, uncovering additional perspectives that may not have been observed during the earlier analytical stages.

## 4. Analyzing News Aggregation on the Israel–Lebanon Conflict: A Comparative Study of Google News

In order to illustrate and validate the presented methodological approach, this study conducts an analysis of data from Google News. Google News is a news aggregation platform that dynamically and continuously compiles headlines from various information sources, organizing them and serving as an amplifier of content from recognized and accredited media outlets [87]. To achieve this, a comparative analysis is performed on the news aggregated by the main pages of Google News Israel (in Hebrew) and Google News Lebanon (in Arabic). The period considered spans thirty days, covering the entire month of November 2024, a timeframe marked by escalating tensions between Israel and Hezbollah, alongside Lebanon's involvement in the broader context of conflicts in the Middle East. The relevance of this study lies in its examination of how a digital news aggregation platform reflects journalistic coverage of topics related to both countries, taking into account linguistic and editorial particularities. This case study aims to provide a descriptive perspective that aids in understanding the reported events, shedding light on the dynamics of narrative construction and its impact on information consumption in crisis scenarios.

The geopolitical tensions between Israel and Lebanon have deep historical roots shaped by territorial disputes, religious divergences, and ideological conflicts [88]. Israel's occupation of southern Lebanon, Hezbollah's influence, and Iran's support for militant groups in the region have intensified these tensions over the past decades. Moreover, recent armed clashes, internal political instability in Lebanon, and the complex regional dynamics, exacerbated by foreign interference, have heightened a fragile landscape [89,90]. In this context, the crisis between Israel and Lebanon escalated significantly in 2023 and 2024, marked by military attacks between the Israel Defense Forces and the Hezbollah paramilitary group, further aggravated by the deaths of Hezbollah leaders such as Hassan Nasrallah, Fu'ad Shukr, and Ibrahim Aqil [91]. Additionally, it is crucial to note that the civilian populations of both countries have been heavily impacted, whether through mass displacement or a high number of casualties [92]. This entire panorama is reflected in both local and international media and, consequently, projected onto news aggregation platforms.

Google News is a news aggregation platform that indexes, organizes, and dynamically makes journalistic content available, presenting continuous variations in temporal and regional terms [87]. Aggregated news is organized through algorithmic curation that groups related content into a uniform structure, with each news item including standard elements such as the source, title, and publication date on the original source. Additionally, the platform is structured with a main page highlighting the most relevant news, complemented by specific thematic sections. An important feature is the platform's ability to offer personalization based on the user's interests and search history. However, in anonymized access scenarios, this personalization does not apply, and the content presented is generic.

Google News does not provide a native API for accessing its data. Therefore, data collection was conducted through the development of a data extraction solution (web scraping) implemented using the Selenium and BeautifulSoup libraries of the Python programming language. This solution directly accesses the platform's webpage and, based on mapping the site's HTML pattern, extracts the variables of interest. It is important to note that the intrinsic features of Google News as a news aggregation platform do not include the storage of news sets delivered at specific moments, nor does it offer a pagination structure. This implies that it is not possible to retrieve retrospective information about the news presented at a given time. To capture a "snapshot" of the news during a specific time interval, a continuous data extraction process must be carried out, even if segmented into regular intervals. Considering the analyzed temporal period (November 2024), the

extraction solution was configured to perform scheduled daily accesses to the Israeli and Lebanese versions of Google News, collecting data from the main page of each version.

The data obtained through the application of the web scraping solution resulted in 953 aggregated news records for the Israeli version of Google News and 990 records for the Lebanese version. The following attributes were considered in the data extraction procedures: aggregation date, source date, source name (news outlet), headline, news url, and Google News version. Figure 5 illustrates the variation in volume over the collection period, where it can be observed that the Lebanese version exhibited no quantitative fluctuations, while the Israeli version displayed variable behavior, with daily volumes ranging from 29 to 34 news items. This slight difference in volume can be attributed to various factors, such as the amount of content generated by local media outlets, the frequency of updates made by the platform, or differences in editorial policies influencing Google News' news aggregation algorithms in each region. Additionally, the geopolitical context and the intensity of newsworthy events may play a significant role, particularly in the case of the Israeli version, where coverage of conflict-related events may be more dynamic and sensitive to daily changes.

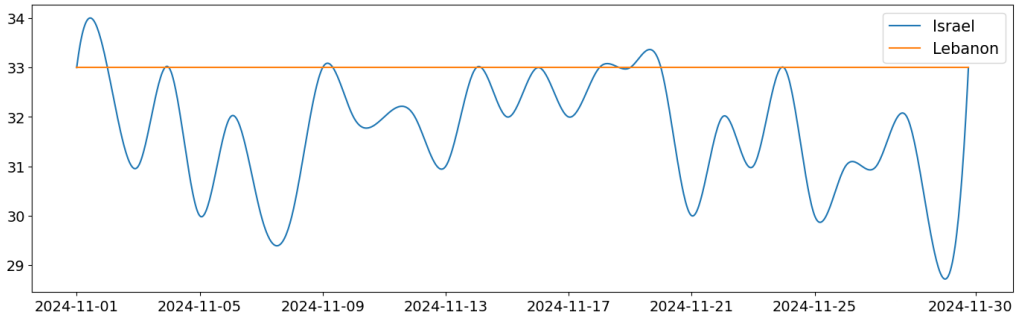



**Figure 5.** Daily volume of news aggregated by version on the Homepage.

The collection of news sources was carried out directly during the execution of the web scraping script, with automated identification of the field corresponding to the source in each record. The analysis of the aggregated news sources allows for mapping the diversity of sources within the platform and aids in understanding how the informational ecosystems of each region reflect local dynamics of news production and dissemination. In the case of Israel, 66 distinct sources were identified, whereas in Lebanon, the number was significantly higher, with 120 different sources. This discrepancy suggests potential structural differences in the media ecosystems of the two countries, which could be related to the number of active media outlets, the variety of sources included by Google News' algorithms in each region, as well as linguistic considerations. For instance, Hebrew is an official language exclusively in Israel, whereas Arabic is spoken across several countries.

Table 1 presents the distribution of news headlines among the five leading sources for each analyzed version, highlighting characteristics related to the representativeness of the media in each context. In Israel, the most prominent sources include **ynet** ידיעות אחרונות (Ynet Yedioth Ahronoth), accounting for 19.94% of the aggregated news, followed by mako with 12.07%, and מעריב און لין (Maariv Online) with 8.39%. These sources represent traditional and well-established media outlets, encompassing both portals associated with major media conglomerates and those with historical relevance. In contrast, in Lebanon, the three most representative sources include الجزيرة نت (Al Jazeera Net), accounting for 7.37% of the news; ليبانون ديبايت (Lebanon Debate), with 6.67%; and Elnashra, with 4.95%. This set reflects a combination of traditional media, such as Al Jazeera, and alternative outlets, like Lebanon Debate, known for its digital and independent focus.

**Table 1.** Distribution of headlines by sources.

| Google News Israel | | Google News Lebanon | |
|---|---|---|---|
| **ynet ידיעות אחרונות** (Ynet Yedioth Ahronoth) | 190 (19.94%) | **الجزيرة نت** (Al Jazeera Net) | 73 (7.37%) |
| **mako** (mako) | 115 (12.07%) | **ليبانون ديبايت** (Lebanon Debate) | 66 (6.67%) |
| **מעריב און לין** (Maariv Online) | 80 (8.39%) | **Elnashra** | 49 (4.95%) |
| **הארץ** (The Country) | 40 (4.20%) | **لبنان ٢٤** (Lebanon 24) | 39 (3.94) |
| **סרגים** (Knitted) | 39 (4.09%) | **القوات اللبنانية** (Lebanese Forces) | 36 (3.64%) |

The presence of **القوات اللبنانية** (Lebanese Forces) among the leading sources in Lebanon warrants attention, as it is a news platform associated with a political party. Its prominence on Google News could suggest that political sources hold significant space within the Lebanese media ecosystem, potentially exacerbating the polarization of narratives surrounding regional conflicts. In contrast, the main Israeli sources listed in the table appear to be less associated with specific political parties, indicating a media structure more focused on commercial conglomerates and generalist portals.

Aligned with the methodological procedures for data enrichment, this case study did not consider derived textual markers, meaning those directly obtained through regular expression-based procedures. On the other hand, to enhance analyses of content circulation based on Google News' news aggregation, the following contextual semantic entities were considered: the names of individuals mentioned in the news titles and the geographic location (country) referenced by the news. In this context, given the linguistic variations and the need for standardization to enable the use of specific NLP solutions and to ensure better comprehension of the generated indicators, the attributes related to news titles and sources were subjected to an automatic translation process into English during preprocessing.

The analysis of mentions of individuals' names in the aggregated news was conducted using a data enrichment solution based on the application of a named entity recognition (NER) model supported by a pre-trained BERT model. For this purpose, the dslim/bert-base-NER implementation was used. The choice of bert-base-NER in this case study was made based on the balance between computational requirements, performance, and accuracy. During the post-processing stage, variations of names referring to the same individual, such as "Netanyahu" and "Benjamin Netanyahu", were consolidated with the support of an empirically created association dictionary derived from observations in the analyzed corpus [13]. This refinement strategy ensured greater uniformity in the analysis and achieved a precision of 95.40% during validation.

Table 2 presents the five most frequently identified names in both versions, showing the numerical total and the percentage relative to the volume of identified names. The results underscore the prominence of political and public figures associated with the analyzed geopolitical context. For Google News Israel, Prime Minister Benjamin Netanyahu was the most mentioned name, with 51 mentions (9.78%). Other notable names include Eli Feldstein (22 mentions), whose presence is linked to leaked confidential documents from the Prime Minister's Office, and Israel Katz (13 mentions), the current Minister of Defense of Israel. These figures were among the most frequently cited in the Israeli version of Google News. Additionally, the names of Donald Trump, the president-elect in the 2024 elections, and Joe Biden, the outgoing president, were also frequently mentioned, highlighting the relevance of U.S. foreign policy in the Israeli context.

Regarding Table 2, in Lebanon, the most frequently mentioned name was Donald Trump, with 49 mentions (9.40%), followed by Benjamin Netanyahu (30 mentions, 7.75%) and Amos Hochstein (19 mentions, 4.90%), the U.S. envoy who led diplomatic efforts for a ceasefire. These indicators suggest a trend toward a greater presence of news addressing perspectives external to the country. The name of the Lebanese Parliament Speaker, Nabih Berri, appeared 11 times (2.84%), being the only public figure directly associated with Lebanon to rank among the five most cited names. Lastly, the inclusion of Formula 1 driver Max Verstappen highlights greater thematic diversity in the content aggregated by Google News Lebanon, even within a context dominated by geopolitical issues.

**Table 2.** Five most mentioned names.

| **Google News Israel** | | **Google News Lebanon** | |
|---|---|---|---|
| (Benjamin) Netanyahu | 51 (9.78%) | (Donald) Trump | 37 (9.56%) |
| (Donald) Trump | 49 (9.40%) | (Benjamin) Netanyahu | 30 (7.75%) |
| (Eli) Feldstein | 22 (0.42%) | (Amos) Hochstein | 19 (4.90%) |
| (Israel) Katz | 13 (0.24%) | (Nabih) Berri | 11 (2.84%) |
| (Joe) Biden | 11 (0.21%) | (Max) Verstappen | 10 (2.58%) |

The second contextual semantic entity pertains to geographic localization, determined by identifying the country referenced in each aggregated news article. In the first step, similarly to name identification, a strategy utilizing a pre-trained BERT model combined with NER techniques (bert-base-uncased-city-country-ner) was employed. In the second step, a BERT-based classifier was applied for country identification (bert-multilingual-uncased-geo-countries-headlines). In cases of discrepancies between the two models, the country corresponding to the analyzed Google News version was associated with the article. Similarly, for records where no geographic origin was identified, the article was linked to the country corresponding to the analyzed Google News version. The validation of the method, based on a pre-labeled sample, achieved a precision of 96.50% [13]. It is important to note that this procedure was not intended to classify the geographic location of the news outlet but to identify potential geographic references in the analyzed headlines' content.

The indicators obtained reveal a strong predominance of news related to the internal context of each country, reflecting the local nature of news aggregation by Google News. In the case of Google News Israel, out of the 953 aggregated news headlines, 647 were identified as referencing the country itself, representing 67.89% of the total records. This predominance suggests a pronounced focus on national event coverage, aligned with the intensification of regional conflicts. Similarly, in Google News Lebanon, 637 out of the 990 aggregated news headlines were related to Lebanon, corresponding to 64.34% of the total. This also highlights significant attention to domestic issues, emphasizing the centrality of the country in the circulation of journalistic content.

In addition to news related to the national context, references to other identified countries were analyzed. Table 3 highlights the five most mentioned countries in each version of Google News, providing both absolute values and relative percentages. In Google News Israel, Lebanon was the most frequently cited foreign country, with 72 mentions (23.53%), followed by the United States (46 mentions, 15.03%) and the Netherlands (30 mentions, 9.80%). Conversely, in Google News Lebanon, Israel was the most mentioned country, with 115 references (32.58%), followed by the United States (48 mentions, 13.60%) and Syria (35 mentions, 9.92%).

These results reflect the regional and international impact of the conflict, highlighting the countries directly involved, such as Israel and Lebanon, as well as external actors like the United States, which plays a mediating and strategic role in the region. Furthermore,

the mentions include countries implicated in the conflict, such as Iran and Syria, as well as the Palestinian Territory in the Lebanese version. It is noteworthy that no mentions of the Palestinian Territory were identified in the records collected from Google News Israel. This absence may be related to editorial positions and biases [93,94].

**Table 3.** Five most referenced countries.

| Google News Israel | | Google News Lebanon | |
|---|---|---|---|
| Lebanon | 72 (23.53%) | Israel | 115 (32.58%) |
| USA | 46 (15.03%) | USA | 48 (13.60%) |
| Netherlands | 30 (9.80%) | Syria | 35 (9.92%) |
| Iran | 22 (7.19%) | Palestinian Territory | 24 (6.80%) |
| Syria | 19 (6.21%) | Iran | 18 (5.10%) |

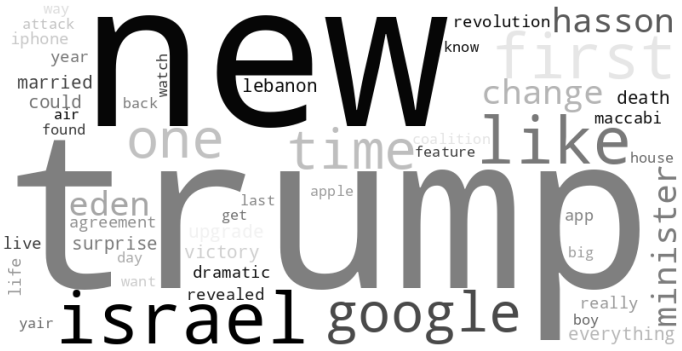
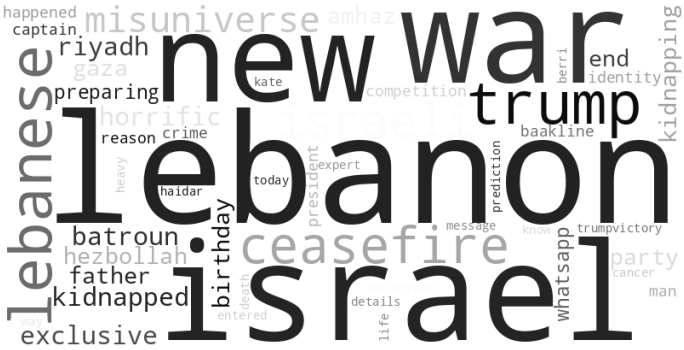
Within the knowledge discovery stage of the proposed methodological framework, and considering the nature of this case study, an analysis of the occurrence of clickbait was conducted to identify the use of sensationalism as a dissemination strategy. For this purpose, a multilingual pre-trained model (xlm-roberta-large) [95] was utilized, achieving an accuracy of 97.59% in the original tests described in the literature [96]. In the case of Google News Israel, 313 records (32.84%) were identified as potential clickbait, while for Google News Lebanon, the identified value was 203 records (20.50%). Table 4 presents the results obtained from the application of the clickbait identification model, consolidated through the generation of a word cloud and a sample selection that exemplifies the headlines classified as potential clickbait. The word clouds presented visualize the frequency of terms within the analyzed corpora, with font sizes proportional to their occurrence (frequency). This prioritization highlights recurring themes in each dataset.

The analysis of clickbait occurrences in Google News Israel records reveals a significant tendency toward the use of sensationalist headlines, mostly unrelated to the geopolitical conflicts involving the country. Frequently used terms such as “Trump”, “new”, “Google”, and “time” indicate a pronounced focus on content related to technology, international political figures, and entertainment. This trend is corroborated by the sample selection, which includes headlines emphasizing technological updates (“Meta gets a major upgrade” and “A small change to Google Authenticator”) and events involving local personalities and celebrities (“For the third time: Einav and Raz got married”), as well as broad statements on changes in public interest issues (“Flights to Israel will be completely banned”). These elements suggest that the use of clickbait in the Israeli version of Google News extends beyond regional security or diplomacy issues, leveraging emotionally appealing or globally popular topics to maximize engagement.

The analysis of Google News Lebanon records, in contrast, shows a predominance of clickbait associated with topics directly related to the geopolitical and military conflicts affecting the country. Frequently used terms such as “Lebanon”, “Israel”, “war”, and “ceasefire” highlight a narrative focus on issues tied to regional tensions, ceasefire negotiations, and the impacts of war. The sample selection reinforces this trend, with headlines blending elements of sensationalism and dramatization to capture attention, as seen in “What does Miss Lebanon say about her participation in the Miss Universe competition during the war?” which combines the emotional appeal of a public figure with the urgency of a wartime context. Additionally, headlines like “Here is the death toll of the Lebanese army since the outbreak of the war” and “Will the ceasefire agreement between Hezbollah and Israel succeed?” appeal to a sense of immediacy and uncertainty regarding conflict developments. This contrast with the more diversified and non-geopolitical focus of Google News Israel reflects, to some extent, the weight that local issues, particularly those related

to war and security, exert on Lebanese media. It also aligns with specific editorial lines, which may vary across different countries.

**Table 4.** Word clouds and sample excerpts for records identified as clickbait.

| **Google News Israel** | |
|---|---|
|  | Meta gets a major upgrade: These are all the new features in the Messenger app ( /Ace)<br><br>A revolution in the military as well: Trump is going to cause a shock in the Holy of Holies of the USA ( /Maariv Online)<br><br>A small change to Google Authenticator will save you some time ( /Geeky)<br><br>For the third time: Einav and Raz from “Chatonami” got married in a large, elaborate event (mako/mako)<br><br>Flights to Israel will be completely banned: This is the dramatic decision announced ( /Ace) |
| **Google News Lebanon** | |
|  | What does Miss Lebanon say about her participation in the Miss Universe competition during the war? (BBC News عربي/BBC News Arabic)<br><br>Here is the death toll of the Lebanese army since the outbreak of the war (LebanonDebate/LebanonDebate)<br><br>After Berri referred to a letter signed by Trump in a restaurant in Dearborn for a ceasefire in Lebanon...this is what the restaurant owner said (LBCI Lebanon/LBCI Lebanon)<br><br>Al-Duwairi: For these reasons, Israel will expand its penetration into Lebanon, and this is what will happen (الجزيرة نت/Al Jazeera Net)<br><br>Will the ceasefire agreement between Hezbollah and Israel succeed? (Lebanon24/Lebanon24) |

Note: the original text was translated into English to facilitate the analyses conducted.

In addition to evaluations based on direct and derived indicators, the knowledge discovery phase incorporated the creation of visualizations that highlight the most frequent terms in the textual corpora formed by inter-references between the countries in the two analyzed versions. Specifically, for Google News Israel, a textual corpus was constructed exclusively from headlines directly referencing Lebanon. Similarly, for the Lebanese version, the textual corpus consisted solely of headlines directly mentioning Israel. Based on these corpora, word clouds and similarity graphs were generated to support the analysis.

Word clouds are visual tools that emphasize the most frequent terms in a textual corpus, with font sizes proportional to their frequency of occurrence [97]. These tools are particularly useful for quickly identifying the most recurring themes, enabling an initial understanding of the focus of headlines in each version of Google News. On the other hand, similarity graphs provide a more structured perspective of relationships between

terms, mapping semantic or co-occurrence connections among words in a visual format. These graphs identify clusters of words that frequently appear together and facilitate an understanding of contextual associations within the analyzed texts [97].

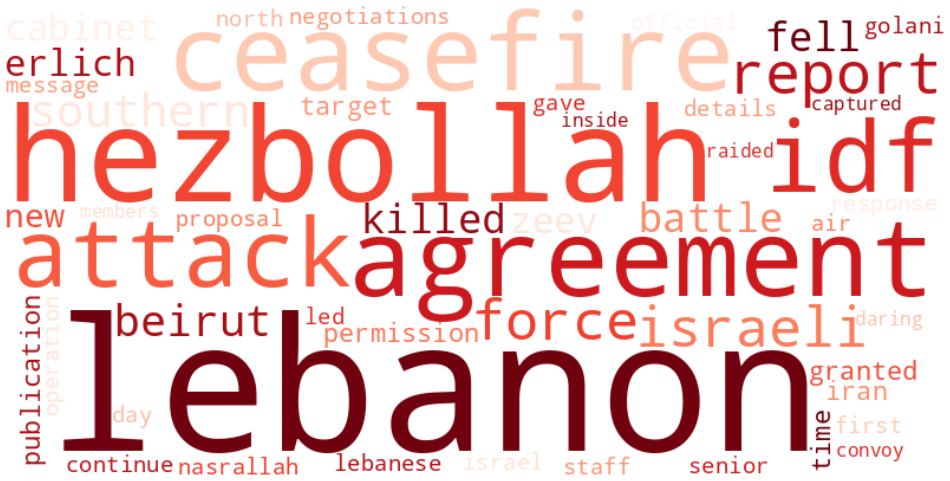
The word cloud generated from the Google News Israel corpus (Figure 6a), which references Lebanon, highlights key terms associated with the Israeli narrative regarding its neighboring country, emphasizing military, geopolitical, and security-related themes. The most frequent terms, such as "Lebanon" and "Hezbollah", underscore the centrality of Lebanon and the political–military group Hezbollah in the analyzed headlines, suggesting a focus on conflict and regional instability. Additionally, words like "attack", "IDF" (Israel Defense Forces), "killed", and "battle" indicate a strong presence of narratives centered on armed confrontations and military actions, while "ceasefire" and "agreement" point to mentions of negotiations and diplomatic efforts. The presence of terms such as "Beirut" and "Southern" highlights specific geographical aspects, while "Iran" reinforces the involvement of external actors in the conflict's context. Similarly, the appearance of figures such as Nasrallah (leader of Hezbollah, assassinated in September 2024) and topics like "report" and "publication" suggest coverage of political events and statements.

Complementarily, the similarity graph generated from the Google News Israel corpus referencing Lebanon illustrates the semantic and contextual connections between the most frequent terms in the analyzed headlines (Figure 6b). In this case study, for the construction of the similarity graphs, the preprocessed textual dataset is subjected to a routine for extracting the most frequent terms, which represent the nodes. Subsequently, the relationships (edges) are established based on the proximity of two terms within a defined context window. Each time two terms co-occur within the same context, a relationship is recorded in a co-occurrence matrix, which ultimately enables the graph's construction. In this study, the graph was rendered using the Gephi visualization software (0.10.0).

In Figure (Figure 6b), the central term "Lebanon" serves as the primary node, with links to other relevant terms, indicating direct associations within journalistic narratives. Terms such as "Hezbollah", "IDF" (Israel Defense Forces), and "attack" exhibit strong connections to "Lebanon", reflecting a focus on the coverage of armed conflicts and military operations involving Hezbollah and Israeli forces. The presence of terms like "ceasefire", "agreement", and "negotiation" suggests the inclusion of topics related to diplomatic efforts and conflict resolution proposals.

Additionally, other nodes such as "Iran", "Golani" (referring to the Golani Brigade of the Israeli Defense Forces), and "Beirut" highlight broader geographical and geopolitical aspects within the context of Israel–Lebanon relations. Terms like "battle", "force", and "operation" reinforce the focus on security dynamics and military engagement. On the other hand, mentions such as "Zeev Erlich" point to the inclusion of specific individuals in the discussion, in this case referring to the death of an archaeologist during a confrontation between the Golani Brigade and Hezbollah combatants [98].

The word cloud presented in Figure 7a corresponds to the textual corpus extracted from Google News Lebanon, which contains mentions of Israel. The most prominent terms, such as "Israel", "Hezbollah", "Lebanon", and "military", reflect their high frequency in the analyzed headlines, indicating the centrality of topics related to the geopolitical conflict between the two countries. The prevalence of words like "missile", "rocket", and "ceasefire" underscores the emphasis on military events, such as attacks and efforts to establish a ceasefire. These aspects are further supported by terms like "agreement", "international", and "court", which suggest mentions of diplomatic efforts, international mediation, or legal repercussions associated with the reported events. Additionally, terms like "Netanyahu" (referring to the Prime Minister of Israel), "Tel Aviv", "Galilee", and "Gaza" point to geographic locations and political figures playing significant roles in this context.




(**a**) Word cloud.


(**b**) Similarity graph.

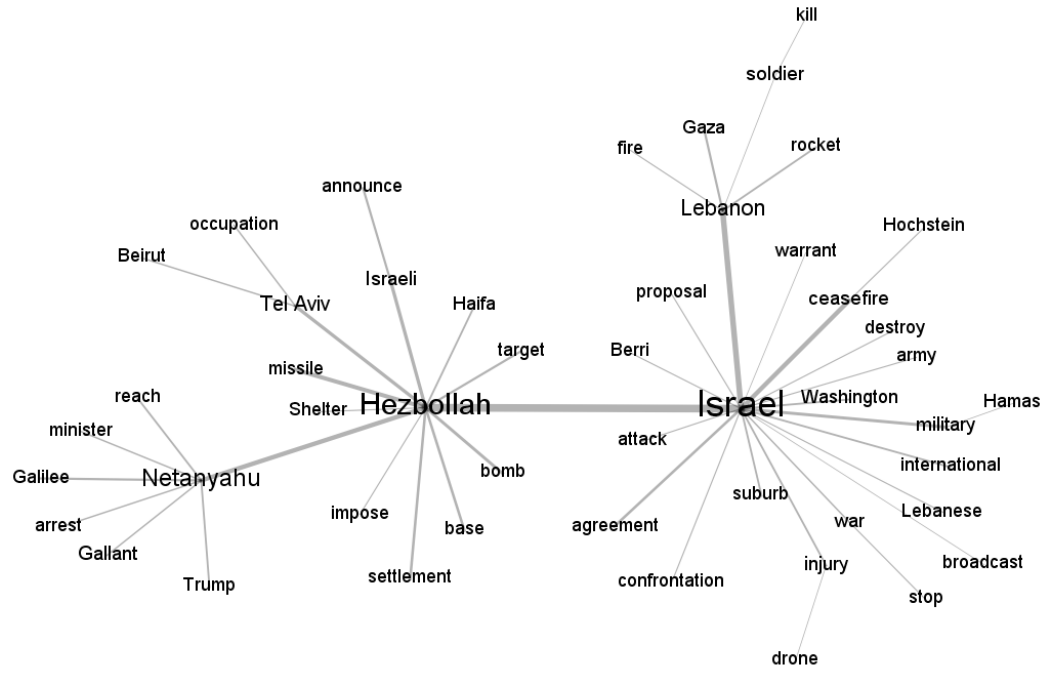
**Figure 6.** Mentions of Lebanon from Google News Israel.

The similarity graph in Figure 7b illustrates the semantic and contextual connections among the main terms from the Google News Lebanon textual corpus referencing Israel. In this graph, "Israel" emerges as a central node with notable connections to terms such as "Lebanon", "Hezbollah", and "Netanyahu", underscoring the centrality of geopolitical, political, and military conflict topics in the region. The robust connection between "Israel" and "Hezbollah" particularly highlights a focus on the coverage of clashes and tensions between Israel and the paramilitary group. Terms like "missile", "target", and "attack" further emphasize the dynamics of the conflict, while the presence of "ceasefire", "agreement", and "proposal" points to diplomatic efforts and mediation attempts aimed at mitigating the confrontation.

Other significant nodes, such as "Haifa" and "Gaza", broaden the geographical scope, highlighting locations also impacted by the conflict. Additionally, the presence of "Netanyahu", "Gallant", and "Trump" reflects the role of political figures in journalistic narratives, signaling discussions on strategic decisions and political implications. The node "Washington", linked to "Israel" and "ceasefire", suggests the involvement of international actors in diplomatic efforts, while "Hamas" and "military" point to additional dimensions

of the conflict involving other groups and military operations. Terms such as "settlement", "occupation", and "international" reinforce the relevance of geopolitical and legal issues frequently associated with territorial disputes.


(**a**) Word cloud.




(**b**) Similarity graph.

**Figure 7.** Mentions of Israel from Google News Lebanon.

The analyses conducted reveal, from a communication perspective, the delineation of narrative construction and discourse reproduction within digital platforms, specifically in the context of news aggregation by Google News. In this regard, the application of the proposed methodological approach in this case study converges toward a scenario that demonstrates the capacity of digital media to shape and amplify public discourse, functioning as arenas where symbolic disputes over the meaning of events are waged. The selection and organization of headlines in news aggregators do not occur neutrally but follow algorithms that prioritize topics and perspectives based on regional consumption patterns and interests [99,100].

The differences in representations between the two analyzed versions, considering all the explored aspects and dimensions, indicate not only the editorial priorities of the aggregators but also the contours of a broader dispute, encompassing the prioritization and circulation of information by journalistic outlets. This process seeks to establish legitimacy, assign responsibility, and shape perceptions [101,102]. In this sense, the results serve both as analytical tools and as resources for developing critical reflections on the role of digital platforms in mediating and amplifying political and geopolitical discourses during crises.

The results achieved through this case study underscore the potential to inform practical strategies for media monitoring and content analysis in real-world scenarios from various perspectives. These include identifying editorial and narrative biases, analyzing real-time trends, understanding communicational and informational dynamics at regional or global levels, defining engagement and communication strategies (including the analysis of clickbait and the circulation of sensationalist content), as well as supporting public policies aimed at transparency and informational diversity.

Finally, it is important to highlight that, although the case study focused on the Israel–Lebanon conflict using data from Google News, the proposed framework has the potential to be adapted to other contexts and media platforms. Scenarios such as the conflicts in Ukraine, the Korean Peninsula, or countries in Latin America could benefit from the described approach, provided adjustments are made to account for the linguistic and cultural specificities of each region. Furthermore, applying the framework to different platforms, such as X or Facebook, would require considerations regarding data collection methods and the specific characteristics of these data.

## 5. Conclusions

Digital media platforms, including news aggregators, social networks, newsletters, blogs, and news portals, have established themselves as essential channels for communication and information dissemination, continuously generating large volumes of textual data. These often unstructured data serve as a significant source of information for research across various fields, enabling the exploration of factors such as narrative analysis, informational flows, and communication strategies. However, the analytical potential of these data is challenged by issues related to collection, processing, and interpretation, particularly in complex scenarios.

In response to these challenges, this study proposed a comprehensive and practical methodology integrating artificial intelligence, statistical methods, and NLP techniques. The methodology covered the entire analysis cycle, from data collection to result visualization, demonstrating a workflow for the application of various techniques and tools. Among its main advantages are the flexibility in data collection, facilitated by the use of web scraping solutions, and data enrichment through advanced natural language processing tools, such as NER and algorithms designed for knowledge discovery. These strategies enable the systematic exploration of large volumes of digital media data, transforming unstructured content into actionable information for specific analyses.

The validation of the proposed methodology was conducted through a case study analyzing news data aggregated by Google News in two regional versions—Israel and Lebanon—during a period of heightened tensions in the Middle East. The results revealed how editorial characteristics, curation algorithms, and geopolitical contexts influence the composition and focus of aggregated news. Furthermore, the analysis highlighted specific communication dynamics, such as the use of clickbait strategies and differences in the representation of events and political actors between the versions analyzed.

Among the limitations of the proposed methodology, it is important to note that data acquisition—whether through APIs or web scraping solutions—is not always guaranteed.

Not all platforms offer accessible or sufficiently comprehensive APIs to meet the needs of complex investigations, and, in many cases, web scraping may be hindered by technical issues, such as frequent changes in webpage structure, or legal restrictions, such as terms of use explicitly prohibiting this practice. Additionally, the proposed methodological strategy relies on artificial intelligence models with sufficient reliability to perform robust analyses. Although the AI models discussed, such as BERT and NER, demonstrated high performance, it is essential to emphasize that they require validation within the specific context of the analyzed data. Thus, the mere application of a model to any given problem without considering contextual specificities is insufficient; it is necessary to verify its applicability, ensure its accuracy is adequate for the dataset, and adjust parameters as needed to guarantee the validity of the inferences made.

As future research directions, the study could be expanded to include comparative analyses across a broader range of algorithmic solutions, contexts, and digital media platforms, enabling, for example, the exploration of cross-cultural dynamics in media representation and consumption. Additionally, the development of automated solutions for adapting to dynamic webpage structures during web scraping would enhance the robustness and scalability of data collection. A deeper exploration of multimodal data—such as combining textual, visual, and audiovisual content—could also provide a more comprehensive understanding of media dynamics in digital environments.

Finally, it is important to emphasize that ethical considerations are essential in studies involving the collection and analysis of digital media data. Compliance with legal and ethical standards must be ensured, particularly concerning user privacy and data ownership. When employing web scraping solutions, it is crucial to respect the terms of service of the platforms being analyzed and to mitigate potential privacy risks.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/fi17020059/s1.

**Author Contributions:** Conceptualization, D.C., C.L., and J.G.; methodology, D.C.; validation, D.C.; investigation, D.C., C.L., and J.G.; data analysis, D.C.; writing—original draft preparation, D.C. and C.L.; writing—review, D.C., C.L., and J.G.; writing—editing, D.C.; visualization, D.C.; supervision, J.G.; project administration, D.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is part of the Project "Parameters and strategies to increase the relevance of media and digital communication in society: curation, visualisation and visibility (CUVICOM)" funded by MICIU/AEI/PID2021-123579OB-I00 and by "ERDF/EU".

**Data Availability Statement:** The data supporting the experiments conducted in the case study are available as Supplementary Material.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Bateman, J.A. What are digital media? *Discourse Context Media* **2021**, *41*, 100502. [CrossRef]
2. Acerbi, A. A cultural evolution approach to digital media. *Front. Hum. Neurosci.* **2016**, *10*, 636. [CrossRef] [PubMed]
3. Moreno, M.A.; D'Angelo, J.D. Digital Media Theory: From One-Way to Multidirectional Communication. In *Handbook of Visual Communication: Theory, Methods, and Media*, 2nd ed.; Josephson, S., Kelly, J., Smith, K., Eds.; Routledge: New York, NY, USA, 2020.
4. Abkenar, S.B.; Kashani, M.H.; Mahdipour, E.; Jameii, S.M. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. *Telemat. Inform.* **2021**, *57*, 101517. [CrossRef] [PubMed]
5. Alotaibi, Y.; Malik, M.N.; Khan, H.H.; Batool, A.; Islam, S.; Alsufyani, A.; Alghamdi, S. Suggestion mining from opinionated text of big social media data. *Comput. Mater. Contin.* **2021**, *68*, 3323–3338. [CrossRef]
6. Labrecque, L.I.; vor dem Esche, J.; Mathwick, C.; Novak, T.P.; Hofacker, C.F. Consumer Power: Evolution in the Digital Age. *J. Interact. Mark.* **2013**, *27*, 257–269. [CrossRef]

7. Mellado, C.; Humanes, M.L.; Scherman, A.; Ovando, A. Do digital platforms really make a difference in content? Mapping journalistic role performance in Chilean print and online news. *Journalism* **2018**, *22*, 358–377. [CrossRef]
8. Mellado, C.; Alfaro, A. Platforms, Journalists and Their Digital Selves. *Digit. J.* **2020**, *8*, 1258–1279. [CrossRef]
9. Subudhi, R.N. Digital Consumption Pattern and Impacts of Social Media: Descriptive Statistical Analysis. In *Trends of Data Science and Applications*; Studies in Computational Intelligence; Rautaray, S., Pemmaraju, P., Mohanty, H., Eds.; Springer: Singapore, 2021; Volume 954, pp. 45–60. [CrossRef]
10. Arrigo, E.; Liberati, C.; Mariani, P. Social Media Data and Users' Preferences: A Statistical Analysis to Support Marketing Communication. *Big Data Res.* **2021**, *24*, 100189. [CrossRef]
11. Vállez, M.; Boté-Vericad, J.J.; Guallar, J.; Bastos, M.T. Indifferent about online traffic: The posting strategies of five news outlets during Musk's acquisition of Twitter. *J. Stud.* **2024**, *25*, 1249–1271. [CrossRef]
12. Men, L.R.; Tsai, W.H.S. Perceptual, attitudinal, and behavioral outcomes of organization–public engagement on corporate social networking sites. *J. Public Relations Res.* **2014**, *26*, 417–435. [CrossRef]
13. Cordeiro, D.F.; Lopezosa, C.; Guallar, J.; Vállez, M. Analysis of Google News coverage: A comparative study of Brazil, Colombia, Mexico, Portugal, and Spain. *Contratexto* **2024**, *42*, 177–208. [CrossRef]
14. Cordeiro, D.F.; Lopezosa, C.; Guallar, J.; Sousa, J.P. News exchange between Brazil and Portugal based on a quantitative analysis of Google News. *Braz. J. Res.* **2025**, *in press*.
15. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37–54. [CrossRef]
16. Shearer, C. The CRISP-DM model: The new blueprint for data mining. *J. Data Warehous.* **2000**, *5*, 13–22.
17. Karlim, Y.P.S.; Hermawan, A.; Maranto, A.R.K. Clustering Mental Health on Instagram Users Using K-Means Algorithm. *Bit-Tech* **2023**, *6*, 32–39. [CrossRef]
18. Viera, L.M.; Cordeiro, D.F. The dark side of anti-vaccination: Analysis of a Brazilian anti-vaccine Facebook group. *Famecos* **2023**, *30*, 1–18. [CrossRef]
19. Garcia-Arteaga, J.; Zambrano-Zambrano, J.; Parraga-Alava, J.; Rodas-Silva, J. An effective approach for identifying keywords as high-quality filters to get emergency-implicated Twitter Spanish data. *Comput. Speech Lang.* **2024**, *84*, 101579. [CrossRef]
20. Ozbay, F.A.; Alatas, B. Fake news detection within online social media using supervised artificial intelligence algorithms. *Phys. A Stat. Mech. Its Appl.* **2020**, *540*, 123174. [CrossRef]
21. Madani, Y.; Erritali, M.; Bouikhalene, B. Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. *Results Phys.* **2021**, *25*, 104266. [CrossRef]
22. Chiang, T.H.C.; Liao, C.S.; Wang, W.C. Investigating the Difference of Fake News Source Credibility Recognition between ANN and BERT Algorithms in Artificial Intelligence. *Appl. Sci.* **2022**, *12*, 7725. [CrossRef]
23. Setiawan, R.; Ponnam, V.S.; Sengan, S.; Anam, M.; Subbiah, C.; Phasinam, K.; Ponnusamy, S. Certain investigation of fake news detection from Facebook and Twitter using artificial intelligence approach. *Wirel. Pers. Commun.* **2022**, *127*, 1737–1762. [CrossRef]
24. Pan, X.; Su, Q.; Wei, L.; Guo, L. Research on Ethical Issues and Coping Strategies of Artificial Intelligence Algorithms Recommending News with the Support of Wireless Sensing Technology. *J. Sens.* **2023**, *2023*, 8629849. [CrossRef]
25. Ma, Y.W.; Chen, J.L.; Chen, L.D.; Huang, Y.M. Intelligent Clickbait News Detection System Based on Artificial Intelligence and Feature Engineering. *IEEE Trans. Eng. Manag.* **2024**, *71*, 12509–12518. [CrossRef]
26. Tu, M. Named entity recognition and emotional viewpoint monitoring in online news using artificial intelligence. *PeerJ Comput. Sci.* **2024**, *10*, e1715. [CrossRef] [PubMed]
27. Guangce, R.; Lei, X. Knowledge discovery of news text based on artificial intelligence. *ACM Trans. Asian -Low-Resour. Lang. Inf. Process. (TALLIP)* **2020**, *20*, 1–18. [CrossRef]
28. Xie, Z.; Wang, J. An artificial intelligence based news feature mining system based on the Internet of Things and multi-sensor fusion. *PeerJ Comput. Sci.* **2023**, *9*, e1428. [CrossRef] [PubMed]
29. Garvey, C.; Maskal, C. Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. *Omics J. Integr. Biol.* **2020**, *24*, 286–299. [CrossRef] [PubMed]
30. Li, J.; Zheng, C. Emotion Classification Method of Financial News Based on Artificial Intelligence. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 8047582. [CrossRef]
31. Jünger, J. Scraping social media data as platform research: A data hermeneutical perspective. In *Challenges and Perspectives of Hate Speech Research*; Strippel, C., Paasch-Colberg, S., Emmer, M., Trebbe, J., Eds.; DEU: Berlin, Germany, 2023; pp. 427–441. [CrossRef]
32. Chani, T.; Olugbara, O.; Mutanga, B. The Problem of Data Extraction in Social Media: A Theoretical Framework. *J. Inf. Syst. Inform.* **2023**, *5*, 1363–1384. [CrossRef]
33. Davidson, B.I.; Wischerath, D.; Racek, D.; Parry, D.A.; Godwin, E.; Hinds, J.; Linden, D.; Roscoe, J.F.; Ayravainen, L.; Cork, A.G. Platform-controlled social media APIs threaten open science. *Nat. Hum. Behav.* **2023**, *7*, 2054–2057. [CrossRef] [PubMed]

34. Trans, M.; Beraldo, D.; Draisci, L.; Afsahi, L.; Brennan, M.; Goldschmidt, V.; Grendarova, T.; Hamilton, H.; Keskin, M.; Papasokratous, M.; et al. APIcalypse Now: Redefining Data Access Regimes in the Face of the Digital Services Act, 2024. Available online: https://www.digitalmethods.net/Dmi/WinterSchool2024APIcalypse (accessed on 30 November 2024). *Digital Methods Initiative Winter School* 2024.
35. Bruns, A. After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Inf. Commun. Soc.* **2019**, *22*, 1544–1566. [CrossRef]
36. Pearson, G.D.H.; Silver, N.A.; Robinson, J.Y.; Azadi, M.; Schillo, B.A.; Kreslake, J.M. Beyond the margin of error: A systematic and replicable audit of the TikTok research API. *Inf. Commun. Soc.* **2024**, 1–19 . [CrossRef]
37. Zhao, B. Web Scraping. In *Encyclopedia of Big Data*; Schintler, L., McNeely, C., Eds.; Springer: Cham, Switzerland, 2022; pp. 345–349. [CrossRef]
38. Mitchell, R. *Web Scraping with Python: Data Extraction from the Modern Web*, 3rd ed.; O'Reilly Media: Sebastopol, CA, USA, 2024.
39. Dewi, L.C.; Meiliana, M.; Chandra, A. Social media web scraping using social media developers API and Regex. *Procedia Comput. Sci.* **2019**, *157*, 444–449. [CrossRef]
40. Thivaharan, S.; Srivatsun, G.; Sarathambekai, S. A Survey on Python Libraries Used for Social Media Content Scraping. In Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 10–12 September 2020; pp. 361–366. [CrossRef]
41. Martín-Gómez, L.; Cordero-Gutiérrez, R.; Pérez-Marcos, J. Business Benefits of Instagram Scraping: Questionable Uses of Data. In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence. DiTTEt 2021. Advances in Intelligent Systems and Computing*; de Paz Santana, J.F., de la Iglesia, D.H., López Rivero, A.J., Eds.; Springer: Cham, Switzerland, 2022; Volume 1410. [CrossRef]
42. Khanom, A.; Kiesow, D.; Zdun, M.; Shyu, C. The news crawler: A big data approach to local information ecosystems. *Media Commun.* **2023**, *11*, 318–329. [CrossRef]
43. Salem, H.; Mazzara, M. Pattern matching-based scraping of news websites. *J. Phys. Conf. Ser.* **2020**, *1694*, 012011. [CrossRef]
44. Barbera, G.; Araujo, L.; Fernandes, S. The Value of Web Data Scraping: An Application to TripAdvisor. *Big Data Cogn. Comput.* **2023**, *7*, 121. [CrossRef]
45. Luscombe, A.; Dick, K.; Walby, K. Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Qual. Quant.* **2022**, *56*, 1023–1044. [CrossRef]
46. Lala, F. Data collection via web scraping: Privacy and facial recognition after Clearview. *I-lex* **2023**, *16*, 34–45. [CrossRef]
47. Fontana, A.G. Web scraping: Jurisprudence and legal doctrines. *J. World Intellect. Prop.* **2024**. [CrossRef]
48. Brow, M.A.; Gruen, A.; Maldoff, G.; Messing, S.; Sanderson, Z.; Zimmer, M. Web scraping for research: Legal, ethical, institutional and scientific considerations. *arXiv* **2024**, arXiv:2410.23432. [CrossRef]
49. Goyal, A.; Gupta, V.; Kumar, M. Recent named entity recognition and classification techniques: A systematic review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]
50. Chew, R.; Wenger, M.; Guillory, J.; Nonnemaker, J.; Kim, A. Identifying electronic nicotine delivery system brands and flavors on Instagram: Natural language processing analysis. *J. Med. Internet Res.* **2022**, *24*, e30257. [CrossRef]
51. Cordeiro, D.F.; Lopezosa, C.; Vállez, M.; Guallar, J. Estrategias de comunicación del Ministerio de Sanidad de España en Instagram antes, durante y después de la pandemia de COVID-19. *Rev. Icono 14 Rev. CientíFica Comun. Tecnol. Emergentes* **2024**, *22*, e2189.
52. Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *Int. J. Mach. Learn. Cybern.* **2024**, 1–65 . [CrossRef]
53. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
54. Tao, L.; Xie, Z.; Xu, D.; Ma, K.; Qiu, Q.; Pan, S.; Huang, B. Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 598. [CrossRef]
55. Object Management Group. Business Process Model and Notation (BPMN), 2013. Available online: https://www.omg.org/spec/BPMN/2.0.2/PDF (accessed on 30 November 2024).
56. Jehangir, B.; Radhakrishnan, S.; Agarwal, R. A survey on named entity recognition - datasets, tools, and methodologies. *Nat. Lang. Process. J.* **2023**, *3*, 100017. [CrossRef]
57. Abzalov, M. Exploratory Data Analysis. In *Applied Mining Geology*; Modern Approaches in Solid Earth Sciences; Springer: Cham, Switzerland, 2016; Volume 12. [CrossRef]
58. Hickman, L.; Thapa, S.; Tay, L.; Cao, M.; Srinivasan, P. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organ. Res. Methods* **2022**, *25*, 114–146. [CrossRef]
59. Anglin, K.L.; Wong, V.C.; Boguslav, A. A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open* **2021**, *7*, 23328584211028615. [CrossRef]
60. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–9 .

61. Mikolov, T.; Corrado, G.; Chen, K.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
62. Mikolov, T.; Yih, W.t.; Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 746–751.
63. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]
64. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
65. Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; Wang, A.H. Twitter spammer detection using data stream clustering. *Inf. Sci.* **2014**, *260*, 64–73. [CrossRef]
66. Li, C.; Liu, M.; Cai, J.; Yu, Y.; Wang, H. Topic Detection and Tracking Based on Windowed DBSCAN and Parallel KNN. *IEEE Access* **2021**, *9*, 3858–3870. [CrossRef]
67. Hanifa, A.; Debora, C.; Hasani, M.F.; Wicaksono, P. Analyzing Views on Presidential Candidates for Election 2024 Based on the Instagram and X Platforms with Text Clustering. *Procedia Comput. Sci.* **2024**, *245*, 730–739. [CrossRef]
68. Mehta, V.; Bawa, S.; Singh, J. WEClustering: Word embeddings based text clustering technique for large datasets. *Complex Intell. Syst.* **2021**, *7*, 3211–3224. [CrossRef]
69. Aurpa, T.T.; Sadik, R.; Ahmed, M.S. Abusive Bangla comments detection on Facebook using transformer-based deep learning models. *Soc. Netw. Anal. Min.* **2021**, *12*, 24. [CrossRef]
70. Singh, J.; Tripathi, P. Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm. In Proceedings of the 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 18–19 June 2021; pp. 193–198. [CrossRef]
71. Kanahuati-Ceballos, M.; Valdivia, L.J. Detection of depressive comments on social media using RNN, LSTM, and random forest: Comparison and optimization. *Soc. Netw. Anal. Min.* **2024**, *14*, 44. [CrossRef]
72. Li, X.; Lei, Y.; Ji, S. BERT- and BiLSTM-Based Sentiment Analysis of Online Chinese Buzzwords. *Future Internet* **2022**, *14*, 332. [CrossRef]
73. Hu, C.; Liu, B.; Ye, Y.; Li, X. Fine-grained classification of drug trafficking based on Instagram hashtags. *Decis. Support Syst.* **2023**, *165*, 113896. [CrossRef]
74. Singh, P.; Jain, B.; Sinha, K. Evaluating Bert and GPT-2 Models for Personalised Linkedin Post Recommendation. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; pp. 1–7. [CrossRef]
75. Jain, B.; Goyal, G.; Sharma, M. Evaluating Emotional Detection & Classification Capabilities of GPT-2 & GPT-Neo Using Textual Data. In Proceedings of the 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 18–19 January 2024; pp. 12–18. [CrossRef]
76. Cui, J.; Wang, Z.; Ho, S.; Cambria, E. Survey on sentiment analysis: Evolution of research methods and topics. *Artif. Intell. Rev.* **2023**, *56*, 8469–8510. [CrossRef] [PubMed]
77. Tan, K.L.; Lee, C.P.; Lim, K.M. A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Appl. Sci.* **2023**, *13*, 4550. [CrossRef]
78. Alslaity, A.; Orji, R. Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behav. Inf. Technol.* **2022**, *43*, 139–164. [CrossRef]
79. Arfat, Y.; Tista, S.C. Bangla Misleading Clickbait Detection Using Ensemble Learning Approach. In Proceedings of the 2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT), Dhaka, Bangladesh, 2–4 May 2024; pp. 184–189. [CrossRef]
80. Hu, B.; Mao, Z.; Zhang, Y. An overview of fake news detection: From a new perspective. *Fundam. Res.* **2024**, *5*, 332–346. [CrossRef]
81. Panda, I.; Singh, J.; Pradhan, G.; Kumari, K. A deep learning framework for clickbait spoiler generation and type identification. *J. Comput. Soc. Sci.* **2024**, *7*, 671–693. [CrossRef]
82. Raj, R.; Sharma, C.; Uttara, R.; Animon, C.R. A literature review on clickbait detection techniques for social media. In Proceedings of the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 14–15 March 2024; pp. 1–5. [CrossRef]
83. Garewal, I.K.; Jha, S.; Mahamuni, C.V. Topic Modeling for Identifying Emerging Trends on Instagram Using Latent Dirichlet Allocation and Non-Negative Matrix Factorization. In Proceedings of the 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 14–15 March 2024; pp. 1103–1110. [CrossRef]

84. Shin, S.H.; Baek, O.J. A Study on Internet News for Patient Safety Campaigns: Focusing on Text Network Analysis and Topic Modeling. *Healthcare* **2024**, *12*, 1914. [CrossRef]
85. Miyers, D.; Mohawesh, R.; Chellaboina, V.; Sathvik, A.L.; Venkatesh, P.; Ho, Y.H.; Henshaw, H.; Alhawawreh, M.; Berdik, D.; Jararweh, Y. Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts. *Clust. Comput.* **2023**, *27*, 1–26. [CrossRef]
86. Yang, J.; Hu, X.; Xiao, G.; Shen, Y. A Survey of Knowledge Enhanced Pre-trained Language Models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2024**, *36*, 1413–1430. [CrossRef]
87. Lopezosa, C.; Vállez, M.; Guallar, J. The vision of Google News from the academy: Scoping review. *Doxa Comun. Rev. Interdiscip. Estud. Comun. Cienc. Soc.* **2024**, *38*, 317–332. [CrossRef]
88. Bukhari, S.R.H.; Iqbal, N.; Khan, A.U. Israel's Military Actions in Palestine and Lebanon: A Critical Analysis of Humanitarian, Political, and Strategic Implications. *J. Dev. Soc. Sci.* **2024**, *5*, 577–586. [CrossRef]
89. Harari, M.; Sözen, A. Lebanon and Israel: Natural Resources and Security Interests as Catalysts for Conflict Resolution. In *Conflict Resolution in the Mediterranean: Energy as a Potential Game-Changer*; Sözen, A., Goren, N., Limon, C., Eds.; Diplomeds—The Council for Mediterranean Diplomacy and Friedrich-Ebert-Stiftung (FES): Amsterdam, The Netherlands, 2024; pp. 23–36.
90. Hassan, Z. Regional Geopolitical Conflict and the Fragile State: Foreign Influence and Lebanon's Sovereignty. In *Reconciliation, Heritage and Social Inclusion in the Middle East and North Africa*; AlDajani, I.M.; Leiner, M., Eds.; Springer: Cham, Switzerland, 2022; p. 27. [CrossRef]
91. Hokayem, E. The Death of Nasrallah and the Fate of Lebanon. *Survival* **2024**, *66*, 33–40. [CrossRef]
92. Nuwayhid, I.; Zurayk, H.; Sibai, A.M. Lebanon: A humanitarian crisis in a complex geopolitical context. *Lancet* **2024**, *404*, 2416–2417. [CrossRef] [PubMed]
93. Barari, H.A.; Yacoub, R. Desmascarando o preconceito da mídia e a influência política impeditiva do sionismo religioso no conflito israelense-palestino. *Am. J. Arts Hum. Sci.* **2024**, *3*, 1–11. [CrossRef]
94. Kareem, A.H.; Najm, Y.M. A critical discourse analysis of the biased role of Western Media in Israeli-Palestinian conflict. *J. Lang. Stud.* **2024**, *8*, 200–215. [CrossRef]
95. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451. [CrossRef]
96. Christodoulou, C. XLM-RoBERTa-Multilingual-Clickbait-Detection. Hugging Face, 2024. Available online: https://huggingface.co/christinacdl/XLM_RoBERTa-Multilingual-Clickbait-Detection (accessed on 10 November 2024).
97. Cao, N.; Cui, W. *Introduction to Text Visualization*; Atlantis Press: Amsterdam, The Netherlands, 2016.
98. Uddin, R. Israeli Archaeologist 'Examining Ancient Site' in Lebanon Killed by Hezbollah. 2024. Available online: https://www.middleeasteye.net/news/israel-archaeologist-ancient-site-lebanon-killed-hezbollah (accessed on 30 November 2024).
99. Fischer, S.; Jaidka, K.; Lelkes, Y. Auditing local news presence on Google News. *Nat. Hum. Behav.* **2020**, *4*, 1236–1244. [CrossRef] [PubMed]
100. Cobos, T.L. Origin and weight of news media outlets indexed on Google News: An exploration of the editions from Brazil, Colombia, and Mexico. *Braz. J. Res.* **2021**, *17*, 28–63. [CrossRef]
101. Ward, S. Patriotism and Journalism. In *Handbook of Patriotism*; Sardoč, M., Ed.; Springer: Cham, Switzerland, 2020; pp. 329–340. [CrossRef]
102. Ojala, M. Is the Age of Impartial Journalism Over? The Neutrality Principle and Audience (Dis)trust in Mainstream News. *J. Stud.* **2021**, *22*, 2042–2060. [CrossRef]