DataPoll: A Tool Facilitating Big Data Research in Social Sciences

Antonis Charalampous, Constantinos Djouvas, and Christos Christodoulou

***Abstract*—The computational analysis of big data has revolutionized social science research, offering unprecedented insights into societal behaviors and trends through digital data from online sources. However, existing tools often face limitations such as technical complexity, single-source dependency, and a narrow range of analytical capabilities, hindering accessibility and effectiveness. This article introduces DataPoll, an end-to-end big data analysis platform designed to democratize computational social science research. DataPoll simplifies data collection, analysis, and visualization, making advanced analytics accessible to researchers of diverse expertise. It supports multisource data integration, innovative analytical features, and interactive dashboards for exploratory and comparative analyses. By fostering collaboration and enabling the integration of new data sources and analysis methods, DataPoll represents a significant advancement in the field. A comprehensive case study on the Ukrainian–Russian conflict demonstrates its capabilities, showcasing how DataPoll can yield actionable insights into complex social phenomena. This tool empowers researchers to harness the potential of big data for impactful and inclusive research.**

***Index Terms*—Big data in social sciences, big data systems, computational methods, computational social science, online platforms, social computing.**

I. Introduction

In an era where people produce and consume massive amounts of digital data on a daily basis—according to Forbes, 2.5 quintillion bytes every day[1]—online media have been established as the dominant medium through which all political actors (i.e., politicians, journalists, and citizens) [1] express themselves across different spectrums of policy issues. Thus, the ability to harvest the enormous amount of data constantly generated and subsequently analyze them using computational methods in order to extract latent knowledge, is rapidly becoming an area of great interest and importance across different research areas [2], including social sciences [3].

This shift towards computational methods of analysis has been driven by a multitude of different reasons. First, the sheer increase in the volume of data being produced from different sources has required the use of automated quantitative methods if the objective is to analyze more than a small subset of them [4]. Manual methods of analysis are simply impractical and too costly when dealing with big data. Technological advances have also played a crucial role in making the key components necessary for computational analysis much more accessible and affordable, while at the same time increasing the computer capabilities in terms of storage and processing power [5]. Furthermore, analysis of massive data sets allows us to delve deeper into phenomena at the societal level and study social behavior from entirely new perspectives [6]. This shift has also been enormously helped by a wide variety of new tools, techniques, and open-source software that constantly develop and facilitate the whole process [7]. Finally, recent advances in the field of natural language processing/understanding [8], which allow a far more accurate and nuanced analysis of text, have opened unprecedented opportunities for the interpretation of human communication using computational methods.

In addition to the above, the nature of digital data plays a significant role in this increase, since digital data exhibit some unique characteristics. According to Davidowitz [9], digital data "do not lie"and people tend to be more honest when expressing themselves in an online setting, since they are not affected by social or environmental interventions, e.g., social desirability. They are also characterized by their pluralism and freshness, since social actors promptly express themselves on different issues in a timely manner.

Recognizing the valuable and unique characteristics of digital data, along with the advances in different technologies that facilitate their analysis, we are currently experiencing a vast growth in social science studies that use them in an attempt to answer various sociocultural questions. Ranging from election forecasting [10], public opinion mining [11], [12], SDGs analysis [13] and fake news detection [14], [15], these studies show how digital data analysis can be used to shed light on different sociocultural phenomena.

At the same time, the emergence of computational social science as an established field with the aim of studying society through the application of computational methods in big data [16] has only served to exacerbate the need for tools that take

Received 7 February 2024; revised 2 August 2024, 17 September 2024, and 11 November 2024; accepted 18 November 2024. *(Corresponding author: Constantinos Djouvas.)*

The authors are with the Department of Communication and Internet Studies, Cyprus University of Technology, Limassol 3036, Cyprus (e-mail: ag.charalampous@edu.cut.ac.cy; costas.tziouvas@cut.ac.cy; ci.xristodoulou@edu.cut.ac.cy).

Digital Object Identifier 10.1109/TCSS.2024.3506582

[1]https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=11dda2a260ba

advantage of these developments and make the field accessible to different groups of people of different expertise by minimizing the technological skills required. Such solutions can prove pivotal in opening exciting new avenues for social science research.

Computational analysis of big data research applied in large text corpora involves three major stages [17]: 1) data collection; 2) data analysis; and 3) data visualization. Each of these steps exhibits its unique characteristics and challenges.

During data collection, one must deal with the sheer quantity, poor quality (highly boisterous), and diversity of online data (highly heterogeneous). Thus, online data collection often requires the creation of different tools that automatically collect, clean, and align the data collected so that they can be processed and analyzed in a uniform way.

Following data collection, data analysis is performed, during which a multitude of techniques are employed to extract latent information from the acquired digital data. One of the most common approaches used for this purpose is machine learning, a laborious and time-consuming process that requires experts' knowledge and experience when done from scratch. Last, in data visualization, different representations (usually graphical) are used, allowing researchers to interpret the results in a more intuitive and comprehensive manner.

Despite these inherent difficulties, the techniques that followed this pipeline proved their effectiveness on different occasions, shedding light on aspects of political content analysis that were not possible before, such as inferring policy positions [18], predicting political orientation [19], and exploring the political agenda [20] among others.

However, their usage revealed some additional limitations. First and foremost, the technical expertise required to properly implement the aforementioned pipeline is not always available, while limitations attributed to the event-specific nature of the proposed approaches translate into techniques that cannot be generalized and reused (e.g., topic modeling applied on a specific corpus). Furthermore, results usually present a snapshot in time, not offering any live observatory features and are usually displayed using static charts (online or on paper), not fully taking advantage of Web 2.0 features and, therefore, not offering any interactivity that will in turn facilitate exploratory analysis.

In this article, we present DataPoll, a novel interactive online tool that aims to address the issues outlined above, enabling interested parties to focus exclusively on addressing the "what" (i.e., what insights should be extracted), leaving the "how" entirely outside of the equation. By exploring innovative ideas, intuitive web interfaces, and state-of-the-art NLU techniques, users will be able to collect and process online data and interpret multilayer analyses on vast amounts of information extracted from different online sources. Whereas other projects such as MediaCloud [21] and EventRegistry [22] focus more on automating specific steps of the pipeline (e.g., data collection), or projects such as Penelope [23] which focus more on creating an infrastructure and open APIs, to the best of our knowledge, this is the first attempt to completely automate the entire pipeline providing an "end-to-end" solution, while at the same time increasing and improving each step in the process.

In addition to bridging the gap between social and computer scientists in terms of the techniques that can be used for collecting and analyzing Big (Digital) Data, we contribute to the community through the following.

1) *Multisource Data Analysis*: DataPoll supports the transparent collection and analysis of digital data from multiple sources, e.g., X, YouTube, Reddit, and Web Media.
2) *Exploratory Analysis*: DataPoll facilitates exploratory analysis through an intuitive and interactive all-in-one (i.e., all individual analyses are presented) dashboard.
3) *Comparative Analysis*: DataPoll supports comparative analysis, facilitating the understanding of the different views different sites have on the same event (it will be better demonstrated in the case study section where the views of Ukraine and Russia regarding their conflict is presented).
4) *Community Collaboration*: DataPoll allows users to extend its features and functionality by adding either new data collections or new data analysis techniques.

In conclusion of this section, it is important to note that although there is skepticism around plug-and-play solutions, since it becomes harder to set up good research designs that are tightly connected to established theories [24], we claim that results obtained through DataPoll can provide valuable insights for those interested in domain-specific analysis.

The article is organized as follows: We begin with an overview of the three main components of big data analysis, followed by a review of similar tools and related research. Next, we describe DataPoll's integration of these components and present a case study on the Russian–Ukrainian conflict, illustrating its features. We conclude with current limitations and future development plans.

II. Computational Analysis of Big Social Science Data

As mentioned in the previous section, computational analysis of big data can be divided into three main steps: 1) data collection; 2) data analysis; and 3) data visualization. This section provides background information on the most common methods used in each of these steps.

A. Data Collection

The first step of the big data pipeline is data collection, i.e., obtaining data relevant to the domain under study. In this section, we present three different techniques used for collecting such data. Depending on the research question(s), the following methods can either be used together or as isolated approaches. However, in most cases, using a combination of these methods is preferable since they complement each other.

First, online databases that contain structured information on a particular type or topic, such as the Harvard Dataverse[2] can be used. Obtaining information from such databases is

[2]https://dataverse.harvard.edu/

usually straightforward, as they provide well-defined procedures for doing so. In these cases, the problem of collecting such data is reduced to the problem of identifying the proper search keywords and parameters that will extract the desired data.

The second option to be used is crawlers. In theory, this approach is the most flexible, since crawlers can collect any information appearing on a web page. This approach requires the building of small dedicated programs known as web scrapers. Unfortunately, it also comes with some significant drawbacks that prevent it from becoming the de facto approach for data collection.

1) Dedicated scrapers must be implemented for each web resource.
2) Constant monitoring is required for potential changes and updates of a website's structure.
3) Differences in structure between sections of the same websites.
4) Data collected using crawlers do not have any inherited structure, thus additional processing that transform them into structured data is required.

Finally, an application programming interface (API) can be used. A requester can collect digital data through API calls, providing different parameters often used for data filtering, such as timeframe, location, language, etc. The great advantage of an API is that it facilitates data collection in a structured and consistent manner. The limitation of this approach is the limited coverage, since not all digital data can be extracted using APIs. Furthermore, most API services either require a paid subscription or introduce usage limits, e.g., the total number of requests per day.

A formal definition of a dataset created using the three methods mentioned above can be as follows:

$$\begin{aligned} \mathcal{D} &= \{D_{\mathrm{X}}, D_{\mathrm{Reddit}}, D_{\mathrm{NewsMedia}}, \ldots\} \\ E_{\mathrm{Source}} &= \{e_1, e_2, \ldots, e_n\} \\ E_{\mathrm{combined}} &= \bigcup_{D_{\mathrm{source}} \in \mathcal{D}} E_{\mathrm{source}}. \end{aligned} \tag{1}$$

Beyond the technical challenges involved in constructing a dataset, researchers must also be aware of the various ethical issues surrounding "human subjects"-based research, such as user consent and privacy, and should be mindful of how they use big data to ensure that users remain protected [25].

B. Data Analysis

In the literature, computational approaches used for analyzing text are traditionally referred to as nature language processing (NLP). In this section, we present five NLP techniques extensively used by social scientists to extract insights from text. These are *topic classification*, *opinion mining*, *name entity recognition*, *ideological scaling* and *network analysis*.

1) Topic Classification: Topic classification is used to assign documents to a set of categories [26] and is an extremely popular content analysis technique among political scientists [27]. Depending on whether categories are known a priori or not, two classes of methods exist, unsupervised and supervised methods.

Unsupervised methods, by taking advantage of underlying features of the text, tries to both estimate a set of categories and position documents within those categories (ibid.). Oftentimes researchers cannot have or do not want to have predefined categories and can therefore utilize these methods to adopt a more exploratory approach and attempt to "discover" the different topics discussed in documents.

Currently, the most popular and widely applied model for unsupervised topic classification is latent Dirichlet allocation (LDA). LDA is a generative probabilistic model based on the idea that "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" [28]. Other models built on top of the LDA architecture are also starting to gain traction among the community, such as the correlated topic model (which tries to model correlation between the occurrence of topics), the dynamic topic model, which captures the evolution of topics over time [29], and more prominently the structural topic model (STM) which extends LDA by facilitating the addition of document meta-data into the topic modeling process [30].

After classification, there is an evaluation phase during which the researcher must manually inspect the clusters produced (each representing a particular topic) and assign appropriate labels to each cluster. Furthermore, the number of topics (i.e., clusters to be generated) must be explicitly specified as an input to the algorithm, often resulting in repeated runs, until the optimal number of topics is found.

Supervised machine learning refers to a set of approaches that are trained using predefined (i.e., annotated) examples. Although one could build their own model that is tailored specifically to their problem at hand, since a huge amount of training data is required to achieve an acceptable benchmark in terms of accuracy, this solution is oftentimes out of reach for most researchers. Alternatively, several pre-trained models are available for use out of the box, which are usually trained on billions of tokens and achieve better performance. One drawback with pre-trained models is that you obviously have no control over the type of data used for training, which might not necessarily be suited to your domain of interest. Thankfully, in our domain of interest (i.e., political text), this problem is mitigated by the fact that most models are trained on news articles, a data type quite broad that fits different cases in social research. In this work, some of the classification features offered are performed by pretrained models.

2) Opinion Mining/Sentiment Analysis: Opinion mining, or otherwise known as sentiment analysis, is "the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language" [31]. It is applied in a multitude of domains, such as the analysis of customer reviews to determine customer satisfaction [32], market intelligence [33], and movie sales prediction [34].

There are two main approaches for automatic sentiment extraction. The first approach relies on a lexicon containing a list of adjectives with their corresponding semantic orientation values. This approach is often referred to as the "dictionary" or "lexicon-based" approach. For any given sentence, the adjectives of each word are extracted and assigned a semantic

orientation value using the dictionary scores. Individual scores are in turn aggregated into a single score representing the overall sentiment of the sentence [35].

The second approach is an implementation of a text classification technique. Different pieces of text are labeled, i.e., assigned a sentiment, which are then used by a machine learning model as training examples for learning the underlying representations. The generated model can classify new data into a category (i.e., one of the labels that existed in the training dataset). Different implementations of this approach exist, derived using different datasets and different machine learning techniques. Generally, this class of approaches outperforms dictionary-based sentiment analysis techniques.

3) Named Entity Recognition: According to Lample et al. [36], the most basic and useful method in NLP is to identify and classify some types of information elements, called named entities. Named entities may belong to one or more predefined semantic types such as person, location, organization, etc. The general goal of this process is to help highlight the fundamental concepts and references in the text and can act as a standalone tool or be part of applications, such as question answering [37] and automatic text summarization [38]. The three main approaches used to solve NER problems are: 1) rule-based approaches, which rely on handcrafted rules and do not need annotated data; 2) unsupervised learning approaches, which utilize unsupervised algorithms such as clustering to extract entities from clustered groups; and 3) supervised learning approaches which require annotated data and treat the problem as one of multiclass classification [39].

4) Ideological Scaling: Scaling models are used for the purpose of estimating the positions of political actors on a latent dimension [4] by using word co-occurrences between texts. Wordscores is a supervised scaling method that requires the selection of reference texts to explicitly define these political positions and uses a similarity score (based on the distribution of words) between these reference texts and new ones to estimate their position [40]. Wordscores is greatly dependent on the reference texts selected, which must contain a substantial amount of ideological speech that also express a position using substantially different vocabulary. Wordfish, on the other hand, is an unsupervised scaling technique that assumes a Poisson distribution of word frequencies, and unlike Wordscores, it does not need any anchoring documents to perform the analysis [41]. As has already been implied, evaluation is critical when scaling algorithms are applied in order to confirm that the ideological space has been identified [27].

5) Social Network Analysis: Social network analysis (SNA), according to Scott and Carrington [42], "is (a) a method to analyze the volume and patterns of social relations linking individual actors to each other, and (b) a way of theorizing the social structure and its effects on behavior". SNA is commonly performed using graph theory, a mathematical approach for studying networks of different kinds. Based on this theory, a social network can be described as a graph composed of a set of nodes (or vertices) that represent social entities or objects, and a set of edges (or lines) connecting vertices that represent the social relations between them [43]. Depending on the type of edge, a graph can either be directed or undirected. For example, on X, user *A* can follow user *B*, but user *B* might not necessarily follow user *A*. This is a case of a directed relationship between two nodes. Whereas on Facebook, if user *A* is friends with user *B*, then automatically user *B* is also friends with user *A*. This is a case of an undirected relationship between the two nodes. Furthermore, several graph algorithms are utilized for SNA. For example, centrality is a common approach used to assess the relevance, or structural importance, of a node in the network [44], consisting of different metrics, such as degree, closeness, and betweenness centrality.

Another major area of interest when analyzing social networks is community detection, since certain types of networks can predict community membership, for example, political orientation [45]. More formally, community detection is concerned with dividing a graph into clusters (i.e., groups of nodes), based on the notion that nodes belonging to the same cluster have some sort of relationship, that is, a more dense connection within clusters than between clusters [44]. To address this problem, several algorithms have been developed, such as the Girvan–Newman algorithm, a hierarchical method that progressively removes "weak" edges from smaller connected components [46], and the Louvain algorithm, a heuristic method based on modularity optimization [47]. A basic modulariy function is defined as

$$Q = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)s_i s_j = \frac{1}{4m}\mathbf{s}^T\mathbf{B}\mathbf{s} \tag{2}$$

where m is the total number of edges, A_{ij} is the adjacency matrix, k_i and k_j are the degrees of nodes i and j, s_i and s_j are elements of the vector s that denotes community membership.

In the literature, there are examples of using this approach for topic modeling, as each community generated can be considered a topic [48], [49].

C. Data Visualization

Data visualization refers to a set of tools that can be used to visualize data graphically. Although there are different definitions and categories of data visualization tools in the literature, in this work we use the Iliinsky and Steele definition. According to Iliinsky and Steele [50], there are two categories of data visualization: explanation and exploration, each having different capabilities and serving different purposes.

Explanatory data visualization is used for the visual representation of a deterministic story that someone wants to visually explain to others, for example, a static pie chart showing gender distribution. Generally, these types of visualization are created using a programming language like Python and R.

On the other hand, exploratory data visualization is used when one wants to experiment with the data in order to extract some insights that cannot be identified a priori. In contrast with explanatory data visualization, this is a non-deterministic approach that allows users to follow any path they want in order to experiment and better understand the data; this is why it can be considered as part of the data analysis phase. Exploratory analysis is achieved using an interactive visual tool, in most

cases through an interactive webpage. Tools that can be used for exploratory analysis are Tableau, Google Data Studio, and Kibana, among others.

Finally, it is worth noting that one can also find and use hybrid solutions, such as Shiny R. These hybrid solutions allow for some minimal interaction and exploration of the data. However, any changes in data might require a recompilation of the code.

III. Related Work

In recent years, there has been a growing interest in applying computational methods for large-scale digital data analysis. This trend is especially evident in the field of computational social science [51], where numerous studies in political and social science leverage text analysis techniques on big data to uncover patterns of individual and group behavior. Below, we provide an overview of these studies, highlighting both their increasing prevalence and the need for tools that support researchers in conducting such analyses.

One category of research focuses on stand-alone techniques like sentiment analysis and topic modeling. For instance, Roy et al. [52] developed a machine learning approach using long short-term memory (LSTM) networks to detect hateful sentiment in tweets, demonstrating superior accuracy over similar models. Proksch et al. [53] introduced a multilingual sentiment-based method that employs automated dictionary translations to measure legislative conflict, significantly reducing research costs and enabling comparative legislative analysis. Similarly, Cochrane et al. [54] found that sentiment dictionaries based on word embeddings outperformed other methods for analyzing emotions in political speeches.

In the realm of topic modeling, Hemphill and Schöpke-Gonzalez [55] used LDA to effectively distinguish topics in political tweets, offering a cost-effective approach for studying political communication. Triga et al. [56], [57] utilized STM to analyze media coverage during the 2018 and 2023 Cyprus presidential elections, while George et al. [58] proposed a hybrid framework combining BERT and LDA to enhance topic modeling applications by preserving semantic information.

Other studies explore the combination of multiple computational methods. Walter and Ophir [59] used a novel computational method combining topic modeling and semantic networks for framing news coverage and electoral success. Kligler-Vilenchik et al. [60] used network analysis, topic modeling, and qualitative analysis to investigate discourses surrounding election mobilization on multilingual X data. Hu and Kearney [61] applied machine learning and computational text analysis to study gender discrepancies in political discourse, while Zehring et al. [62] analyzed key topics within a Telegram network related to COVID-19 protests in Germany using network analysis and STM.

Large language models (LLMs) represent another cutting-edge development in the application of computational methods to social science. Ziems et al. [63] evaluated the zero-shot capabilities of 13 open-source and proprietary LLMs across an extensive suite of 25 representative English CSS benchmarks. After a rigorous and thorough experimentation process, the authors conclude that *"LLMs can augment but not entirely replace the traditional CSS research pipeline."* These models certainly have the potential to reshape the field by automating complex tasks at unprecedented scales, complementing traditional methods.

Computational methods are advancing the speed and breadth of discovery in social research [64]. This study aims to contribute by introducing a novel tool that enhances accessibility to these methods through well-defined, reusable workflows.

A. Online Databases and Platforms

In this section, we present five platforms that share some pipeline functionalities with DataPoll. These are: 1) *media cloud*; 2) *event registry*; 3) *GDELT*; 4) *LexisNexis*; and 5) *Penelope*.

Media cloud is an "open-source data corpus and suite of web-based analytics tools" that offers rich datasets around specific queries or topics, available through both an online interface and an open API [21]. Media Cloud primarily focuses on data retrieval, covering various sources organized by themes or source types.

Event registry [22], another popular platform, provides access to articles and events via both a web app and API. Unlike media cloud, event registry goes beyond data retrieval by offering text analytics services such as sentiment detection, categorization, and semantic similarity, making it a more comprehensive tool for content analysis.

GDELT [65] is a large-scale event database featuring over 200 million geolocated events dating back to 1979. GDELT provides web-based tools and an API, emphasizing metadata assignment (e.g., location, language, and tone). While media cloud and event registry focus on thematic areas and sentiment, GDELT excels in long-term, global event data and geospatial analysis.

LexisNexis [66], a proprietary database, offers a curated selection of formal media outlets, with content licensed directly from publishers. Unlike the open-source platforms, LexisNexis is tailored for academic research, covering 200 countries in 37 languages over the past 45 years, though it lacks the flexibility of tools like Media Cloud and Penelope.

Last, Penelope [23] is open-source and specializes in studying cultural or societal conflicts using social media data. Unlike media cloud and event registry, which provide ready-to-use tools, Penelope requires users to develop observatories, demanding programming skills for retrieving and visualizing results. This makes it highly customizable but limited in scope and more technically demanding.

It is worth noting that some of these tools are already integrated into DataPoll via their APIs, reflecting our effort to apply analysis across diverse resources.

The primary focus of the tools and platforms discussed above is broad coverage of various online news media and blogs, with some also offering analytical capabilities. In this work, we introduce a new computational suite, which we believe is better suited for social science research. Rather than wide source coverage, DataPoll emphasizes targeted, cross-platform media

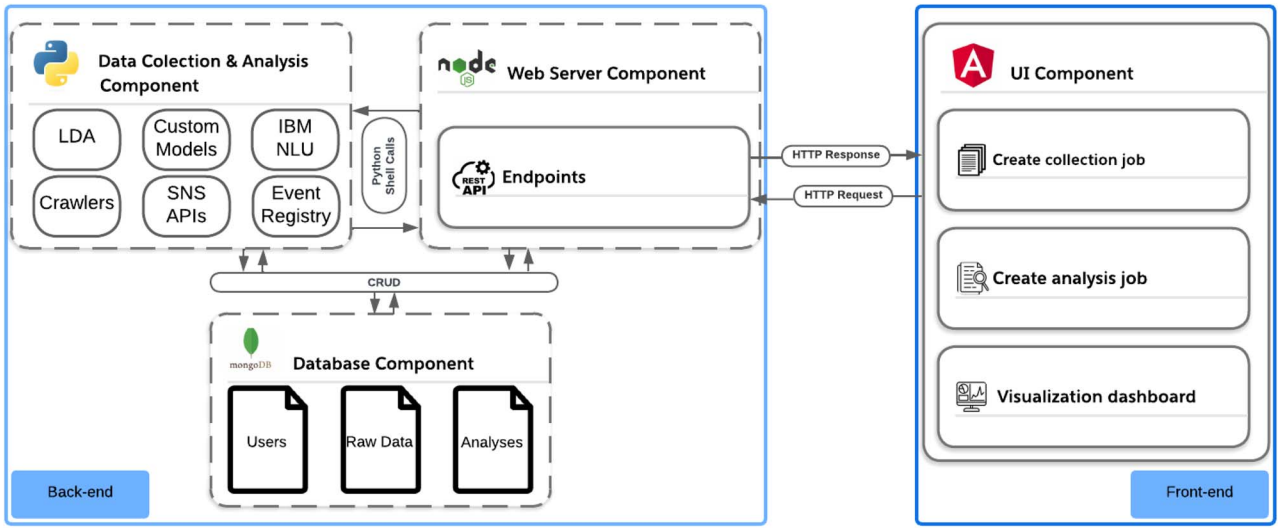

Fig. 1. DataPoll's architecture.

coverage and provides in-depth analysis using a range of computational techniques, along with exploratory and comparative analysis through an interactive dashboard.

IV. DataPoll

This work proposes DataPoll, a tool that aims to simplify, unify, and extend the best practices currently used in Computational Social Sciences. In this section, we first provide a detailed explanation of its features for each of the three major stages in big data research: a) data collection; b) data analysis; and c) data visualization. Then we present DataPoll's its implementation and deployment strategies adopted. An example demonstrating in practice all the features described here is presented in the next section.

A. Architecture

The DataPoll architecture (Fig. 1) consists of two overarching modules: the front-end and the back-end, each containing several submodules that work together to deliver the platform's comprehensive functionalities.

The *front-end* of DataPoll is built using Angular, a robust framework known for creating dynamic and responsive web applications. It serves as the entry point for all the major functionalities available in DataPoll through an intuitive and user-friendly interface. Users can create collection jobs, perform analysis tasks, and interact with the visualization dashboard, built using *ECharts.js* [67]. This ensures that users can efficiently navigate and utilize various functionalities such as data filtering and comparative analysis, among other features.

The *back-end* consists of three main components. The web server, developed using Node.js, is responsible for handling API requests from the front-end, delegating them to one of the other two components of the back-end. It performs necessary actions such as CRUD operations on the database and initiating Python shell processes for data collection and analysis. These processes (e.g., data collection, data analysis, etc.) are executed using dedicated Python scripts running on the background. Python was chosen because of its extensive ecosystem and a rich array of libraries suited for these tasks and especially for NLP. Last, the database is built using MongoDB, a NoSQL database that synergizes well with the rest of the technologies due to its JSON-like document structure.

Since DataPoll follows a microservice pattern, we are able to containerize each microservice using Docker [68] and orchestrate our system with Kubernetes [69], which dynamically scales our services based on CPU and memory usage usage. We utilize Kubernetes' horizontal pod autoscaler to automatically adjust the number of running instances of each service in response to varying load. For memory management, we set resource limits and requests for each container, allowing Kubernetes to efficiently allocate resources and prevent any single service from consuming excessive memory. Last, we implement caching strategies (e.g., Redis) to store frequently accessed data and reduce computational overhead.

Regarding data safety, we use daily backups using a background script that runs daily to backup the database to a secondary server. We can consider improving this method using techniques similar to those described in [70].

B. Data Collection

When it comes to data collection, one of our core objectives is to give researchers the ability to collect data from a multitude of online sources. This emerges from the fact that each online source has its own characteristics, influencing what and how topics are discussed and propagated [71], [72]. Having the capacity to collect data from multiple sources will allow researchers to build a more diverse and complete dataset, leading to a broader and deeper understanding of the sociocultural phenomena they investigate.

Online data sources can be split into two main categories; those that offer a programmatic interface for querying and collecting data, i.e., an API, and those that do not offer such

features, e.g., online news media. Generally, the integration of APIs into DataPoll is simple. Furthermore, credentials required to grant access to an API can be provided to the system and are subsequently encrypted for obvious safety reasons. For traditional news media sources that do not offer APIs, such as online newspapers, a per-case approach must be adopted through dedicated crawlers. End users can create new crawlers using a simple and intuitive web interface for any website they want to crawl.

APIs already integrated in DataPoll include *Twtter*, *YouTube*, *Reddit* and *EventRegistry*. Furthermore, more than 60 preconfigured news media crawlers are already provided and can be used to scrape popular newspaper websites such as *BBC*, *The Guardian* and *Daily Maiil*. These have been implemented using our own *Webmedia* module, built on top of the Scrapy Python library. Through this module, users can also create their own crawlers, eliminating a lot of the boilerplate and offering additional features such as content filtering and automatic generation of CSS selectors.

1) Data Tagging: Tags represent a powerful way to group collections together. A tag can be assigned to multiple collections and a collection can have multiple tags. These two permutations allow for great flexibility in creating and adjusting dataset clusters that can be subsequently treated as a single entity for the next stage of analysis. For example, data regarding the same topic but originating from two different sources can be assigned the same tag in order to designate thematic similarity and group them together.

2) Live Data Monitoring: As mentioned previously, the initial idea of this work was to provide live monitoring of online text, including social networks and news media. Due to the ever-changing nature of online data with new content constantly being generated, it is of paramount importance to perform regular updates and/or augmentations of the data collected.

With this in mind, DataPoll has been designed to facilitate regular and timely collection of data, thus ensuring that the datasets remain up-to-date. This is achieved through the implementation of cron-like schedulers that allow users to set up automated data collection jobs at fixed intervals (e.g., every 2 h, every two days). This feature ensures that researchers can continuously monitor live data streams and capture the most recent information relevant to their studies.

Some of the challenges associated with this process, such as efficient management of large-scale data streams and balancing the computational load to avoid system overloads are handled by the architecture, described in Section IV-A.

3) Data Privacy: DataPoll is committed to ensuring data privacy and adhering to ethical standards, particularly given the sensitive nature of social data. Although the platform allows users to upload their own custom datasets, it is important to note that it is not possible for DataPoll to ensure that all user-uploaded data comply with the required privacy standards. However, DataPoll takes significant measures to protect privacy within the platform. Sensitive information that could potentially identify individuals is not visible in any of the platform's modules, including collection, analysis, and visualization. The charts and visualizations generated by DataPoll present only aggregated results, further safeguarding individual privacy. These measures underscore DataPoll's commitment to ethical data handling and privacy protection.

4) Other Features:

a) Preprocessing: Preprocessing is an important step prior to data analysis with the ability to significantly influence analysis results [73]. Therefore, researchers must be aware of the potential impact of their choices. Currently, users can apply several common preprocessing techniques before analyzing their data. These techniques include: 1) lemmatization; 2) stopword removal; 3) lowering; 4) stemming; and 5) removal of punctuation, numbers, URLs, etc.

b) Translation: Since most machine learning models are trained in a specific set of languages (with by far the most dominant one being English), it is often preferable, more efficient, and more effective to translate the data to one of the supported languages instead of training a classifier from scratch. With that in mind, the tool offers the option of translating text into a different language. This is implemented using the Google translate API; thus, it supports the translation from and to any language supported by Google translate.

c) Custom data upload: Users can easily upload their own data to DataPoll from a CSV or JSON file. Once uploaded, the data will be treated by the system as any other data collection. This allows the application of any feature of the tool on the uploaded dataset, e.g., translation, sentiment analysis, etc.

C. Data Analysis

During this stage, users can apply one or more text analysis techniques to any of their available dataset clusters. DataPoll supports a variety of techniques, including machine learning and advanced statistical models. The platform currently integrates both open-source (e.g., DistilBERT) and proprietary models (e.g. IBM Watson) and implementation is facilitated through a user-friendly interface that allows researchers to select and configure their desired analytical methods without needing extensive programming knowledge. Current techniques implemented in DataPoll include: 1) *sentiment analysis*; 2) *supervised and unsupervised topic classification*; 3) *keyword and concept detection*; 4) *named entity recognition*; 5) *network analysis*; and 6) *statistical measures* (e.g. tf-idf and word frequency)

1) Community Analyzers: To create an environment promoting collaboration between researchers, we provide the option and infrastructure for users to upload their own "analyzers" to DataPoll. The platform provides detailed instructions along with examples on how a user can create and upload a custom analysis module (e.g., advanced statistical models or machine learning models, etc.). The only requirement for the uploaded modules is to follow a specific input and output template, which ensures compatibility and seamless integration with DataPoll's existing framework. By adhering to this template, researchers can extend the platform's capabilities with their specialized methods. We hope that this feature will gradually lead to the development of a rich catalog of innovative analysis techniques, resulting in exciting new insights to be drawn from the data.

D. Data Visualization

The final step of the pipeline is to present the derived results on an interactive dashboard. More precisely, each result is represented as a separate graph (depending on the output of the analysis different graphs can be selected, such as pie charts, bar charts, word clouds, etc.) where various options aimed at facilitating exploratory data visualization (as defined in section data visualization) and multilayer interpretation are made available to the user.

1) Data Manipulation and Hypothesis Testing: DataPoll offers robust interactive features for visualizing and manipulating data, which facilitate dynamic exploration and hypothesis testing. The first major feature, defined as *incremental filtering*, allows users to filter results based on several different conditions. Each additional condition extends all previous ones, thus creating multiple unique representations of the same topic.

Currently, users can filter results on four different axes, outlined as follows.

1) *Date Axis*: Filter results based on a starting and ending date, provided that a date is present on the data analyzed (e.g., show analyses results only of documents from the last three months).
2) *Content Axis*: Filter results based on whether a document contains or does not contain a particular word or phrase (e.g., show analyses results only of documents containing the phrase "COVID vaccine"). This axis can have multiple values.
3) *Collection Axis*: Filter results based on whether they belong to a particular collection or not (e.g., show analysis results only of documents from the "Tweet collection").
4) *Analysis Result Axis*: Filter results based on a specific analysis value (e.g., show analyses results only of documents with a positive sentiment). This axis can have multiple values.

By combining multiple filters on different axes (e.g., date and collection filtering) and/or within the same axis (e.g., sentiment analysis results and topic classification results), we hope to create a highly interactive environment through which end-users can: 1) view an up-to-date graphical representation of the produced results; and 2) apply exploratory analysis.

The second major feature allows users to perform a comparative analysis by displaying two analysis visualizations next to each other. As shown in the following section, we utilized this feature to perform a comparative analysis of the rhetoric used regarding the conflict in Ukraine between Russian and Ukrainian outlets. Different examples of the *data visualization* module are also presented.

These interactive features also support hypothesis testing by allowing researchers to dynamically manipulate and filter the data. For example, if a researcher hypothesizes that documents classified as having "negative" sentiment discuss "war and unrest" more prominently than those classified as "positive," they can test this hypothesis by applying the corresponding sentiment filters and comparing the topic classification results in the visualization dashboard. This flexibility enables users to explore various hypotheses and gain deeper insights through interactive and comparative data analysis.

V. Case Study

This section demonstrates the capabilities of DataPoll through a real example. We decided to use a quite prominent topic of our days, that of the Russian–Ukrainian conflict. This is because, in addition to being a highly polarized topic among the two sides, it also exhibits some unique challenges related to the ban of different Russian news media and social network accounts. However, as shown below, DataPoll can solve this problem using its multisource data collection capabilities. Furthermore, we demonstrate the tools' data analysis capabilities by applying ten different analysis techniques (detailed in the data analysis subsection under the case study section) on the collected dataset clusters. Finally, several figures are included depicting each stage.

A. Data Collection

We started by collecting data from three different sources, X, Reddit, and online news media. We wanted data from each source type to be symmetrical, i.e., generated by an authority or group with similar power and influence; thus we selected the sources accordingly.

For X, we selected the two accounts belonging to the Ministry of Foreign Affairs of each country and used the User timeline lookup endpoint to collect their latest (the API returns up to 3200 tweets), organic (i.e., without retweets) tweets.

For Reddit, we identified two subreddits of interest: r/Ukraine and r/Russia, and proceeded to collect comments from each. In the case of r/Ukraine, we used the official Reddit API for subreddit search and collected the top 300 top-level comments from the ten most popular posts published within the last month, totaling 3000 comments.

In the case of r/Russia, we had to employ a different strategy since Reddit had placed the subreddit in quarantine, making data collection through the official API impossible. To overcome this problem, we used the comment search endpoint of the Reddit Pushshift API and collected approximately 35 000 comments posted in February (the last time point before the subreddit was quarantined). From these, we extracted only the top-level comments from the top 20 submissions (in terms of the number of comments), resulting in a final dataset of 1526 comments. Fig. 2 shows an example of the DataPoll interface for API sources, namely the X user timeline.

Finally, for the news media, we selected one news medium from Ukraine ("Ukrinform"), and two news media from Russia ("Sputnik" and "RT"). For "Ukrinform", we used DataPoll's interface to create a custom crawler that crawled the "War" section of the news website (Fig. 3), collecting 36 news articles. However, applying the same technique on the two Russian news media was impossible due to their ban in the EU. Thus, we implemented a different strategy using the Article search endpoint of the event registry API implemented by DataPoll, collecting 20 articles from each source using the event registry "Ukrainian Crisis" topic. For RT specifically, since the content

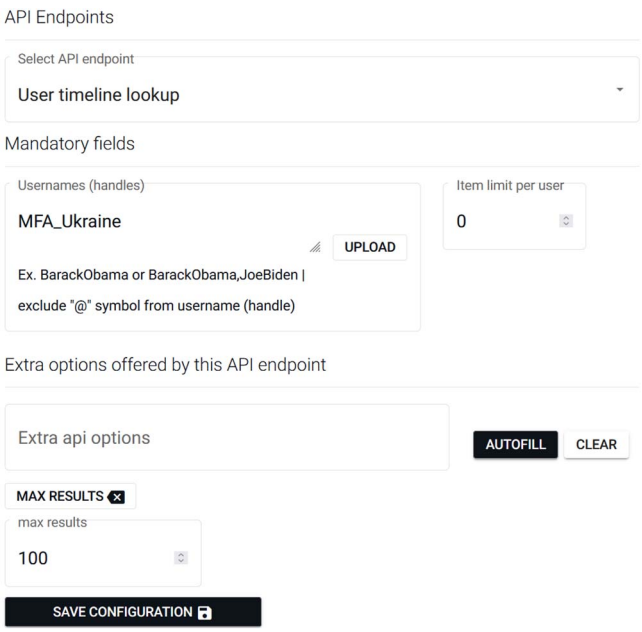

Fig. 2. DataPoll's interface example for API calls.

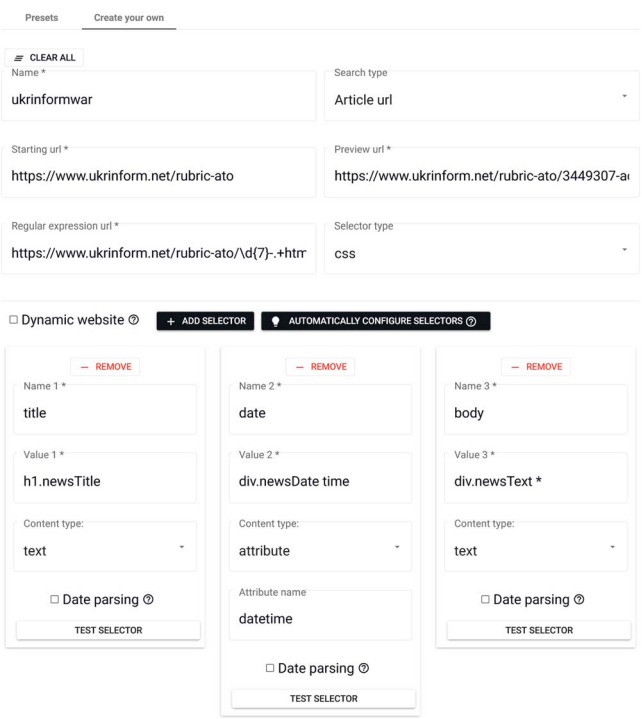

Fig. 3. DataPoll's interface for Webmedia crawler configuration.

was in Russian, we utilized the DataPoll translation feature to translate it to English.

After configuring each collection job, we click "START JOB", initiating the collection and storing of the collected data.

The last figure of this section (Fig. 4) shows DataPoll's page that lists all the data collected, i.e., contain the "Ukraine–Russia" tag. In the next section, all the different collected data will be handled and analyzed as a single unit since all

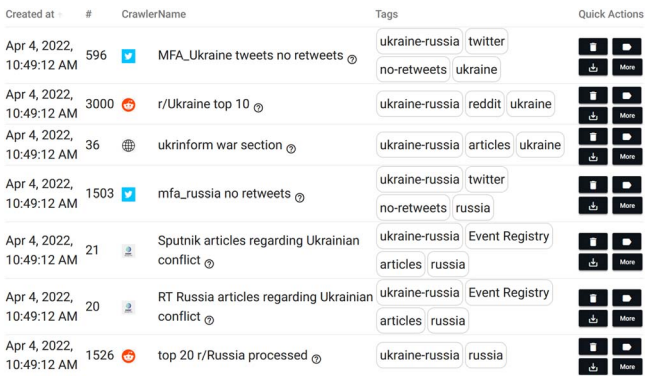

Fig. 4. Final list of collections.

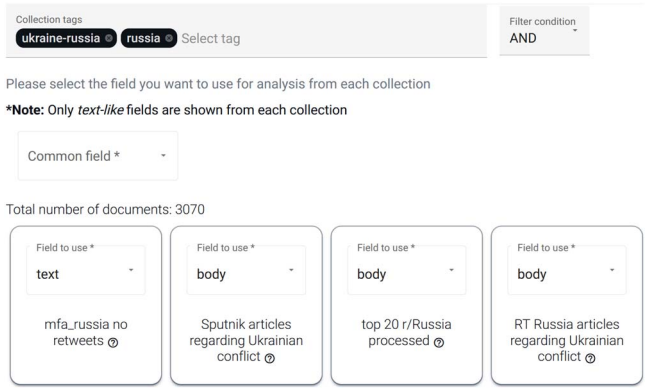

Fig. 5. Datasets for Russian sources filtered using "Ukraine–Russia" and "Russia" tags.

data are automatically aligned and harmonized during the data collection.

For YouTube we were unfortunately unable to collect any data, since Russian media outlets were removed from the platform and we had no alternatives.

In closing this section, we should make clear that the main purpose was not to collect massive amounts of data, but rather to demonstrate the different capabilities of DataPoll regarding data collection. We can, of course, collect more data by selecting additional social network accounts or news media sources; however, a broader data collection process is beyond the scope of this work.

B. Data Analysis

Continuing with the analysis stage, we decided to run two separate analyzes, since we wanted to use comparative analysis to compare the rhetoric used by each side regarding the conflict. To isolate collections belonging to Russian and Ukrainian sources, we used the tag features offered by DataPoll as shown in Figs. 5 and 6.

Last, we had to determine which of the 15 analyzers currently implemented were going to be used to analyze the collected data (the complete list of the analyzers currently implemented is shown in Fig. 7).

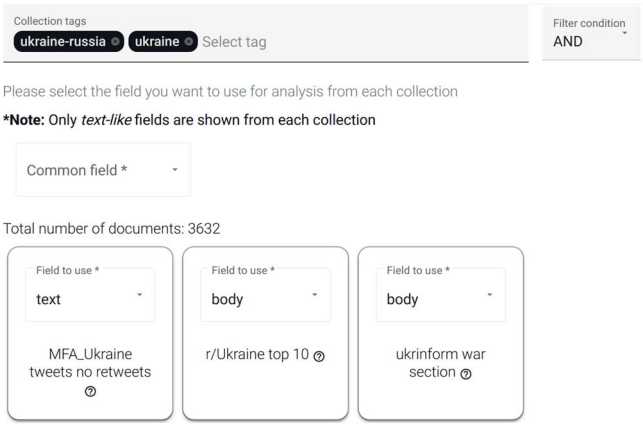

Fig. 6. Collections—Ukrainian sources. Tags used: "Ukraine–Russia" and "Ukraine."

Fig. 7. Final list of analyzers selected (same set of analyzers were selected for both Russian and Ukrainian sources).

We selected the above analyzers (and not all analyzers) because we believe that those are sufficient for demonstrating both the capabilities of DataPoll and also the simplicity of its usage (which is the purpose and this succession), as well as getting some inside information on the different rhetoric used by the two sites.

After configuring each analysis job (i.e., designating the datasets to be analyzed and the analyzers to be applied), one should click the "START JOB" to initiate the analysis and store the results in the database.

C. Data Visualization

This final stage involves exploring and interpreting the analysis results presented through an interactive dashboard. For the purposes of this case study, we only present a subset of the results generated, which were both interesting and also exhibited some of the characteristics described in section *data visualization*. The complete analysis and experiment can be found on the interactive dashboard at the following URL: https://datapoll.app/visualization. However, we must make clear that explaining the results is beyond the scope of this work. This can be the task of a social scientist using the interactive dashboard

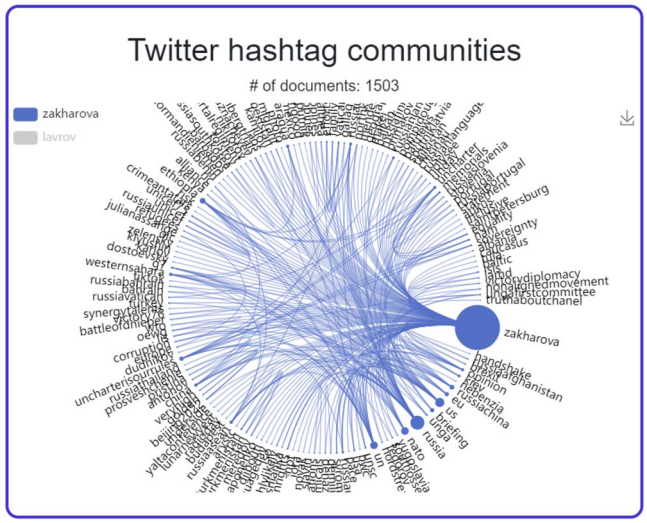

Fig. 8. Russian hashtag community 1.

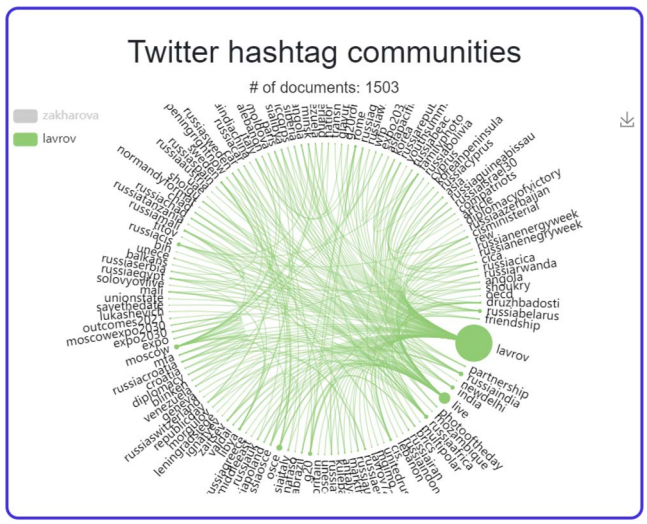

Fig. 9. Russian hashtag community 2.

1) X Hashtag Communities and Network Analysis Results: Looking at the hashtag communities for the Russian data we can clearly see that the most important node in the largest community is "zakharova" with a degree centrality of 0.45 (Fig. 8) while the most important node in the second largest community is "lavrov" with a degree centrality of 0.37 (Fig. 9).

For Ukrainian data, we can observe that the most important node in the largest community is "Ukraine" with a degree centrality of 0.43 (Fig. 10) and similarly "standwithukraine" in the second largest community with a degree centrality of 0.33 (Fig. 11).

These results may give an indication of the different strategies employed by each country's MFA X account. While MFA tweets from Russia seem to place a higher emphasis on foreign

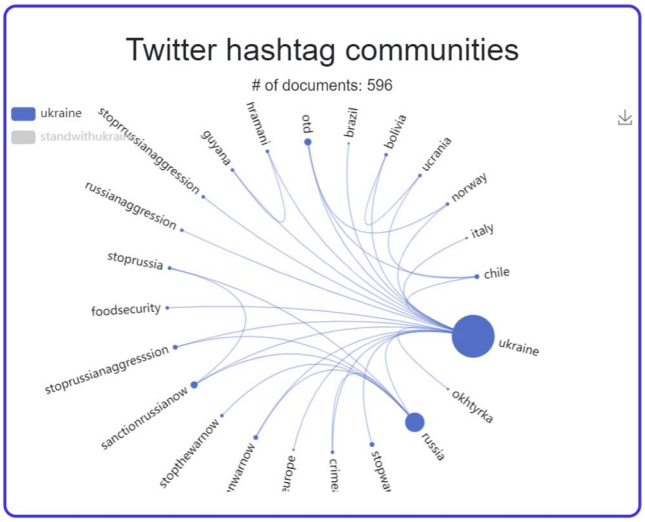

Fig. 10. Ukrainian hashtag community 1.

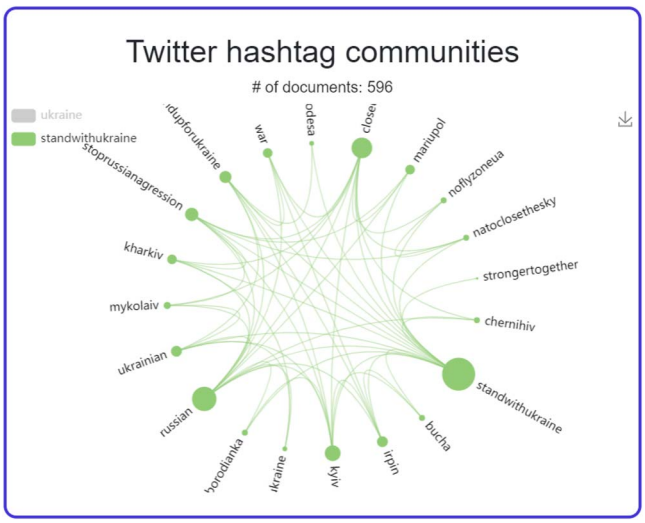

Fig. 11. Ukrainian hashtag community 2.

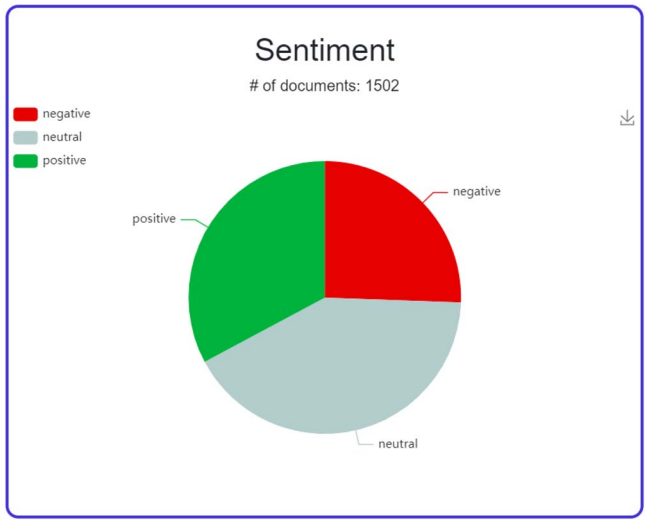

Fig. 12. Sentiment of Russian tweets.

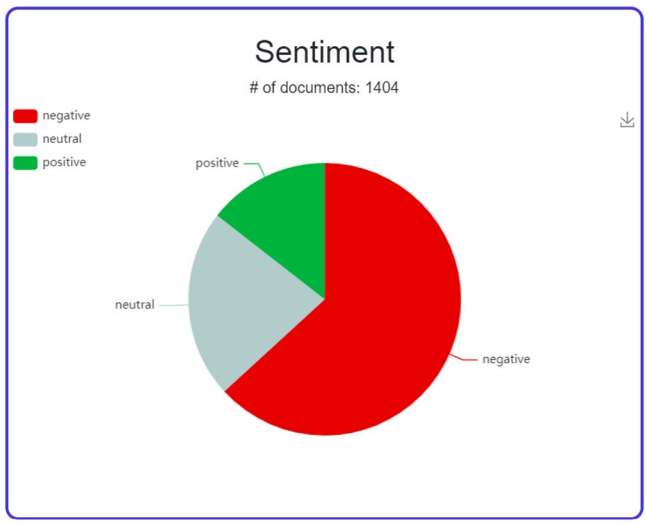

Fig. 13. Sentiment Russian Reddit comments.

policy, tweets from MFA Ukraine appear to focus more on national support.

2) Sentiment Comparison Between Russian Sources: Looking at the sentiment analysis of Russian resources, we can observe a big difference in the level of polarization among the three different sources (X, Reddit, and online newspapers). Whereas for X (Fig. 12) the largest category is neutral with 41.6%, followed by positive with 32.8% and negative with 25.6%, for Reddit (Fig. 13), the neutral category is much smaller at 22.8%, while the negative category is 63% (i.e., 2.5 times larger). For news articles (Fig. 14) the contrast is even more stark, with only 7.5% classified as neutral and 88% classified as negative (73%) or positive (19.5%).

This could be explained by the fact that an official government source (such as the MFA X account) is expected to maintain a more neutral tone, while comments on Reddit made by the general public are more likely to contain polarizing language. News articles, on the other hand, have the most polarized content, possible evidence of controlled media with propagandistic content.

3) Topic Classification Comparison Between Russia and Ukraine: In terms of the topics discussed, we can observe some clear similarities between the two analyses (Figs. 15 and 16). More specifically, the top three categories for each analysis are almost identical. More specifically, "unrest and war" is the top category with 32.7% of documents classified as such in the Russian dataset and 22.7% in the Ukrainian dataset. Next, we have "law, government, and politics," with 7.9% and 4.9%, respectively, while "politics" comes third with 6.3% and 4.5%, respectively. In general, these results suggest that both sites are concerned about the same issues.

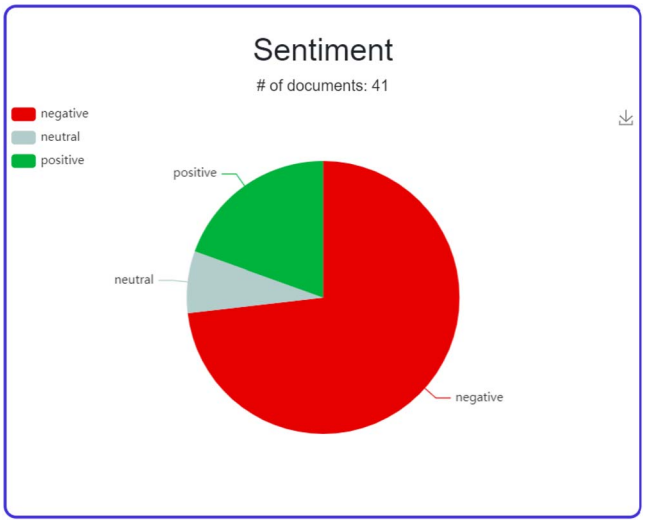

Fig. 14. Sentiment of Russian newspaper articles.

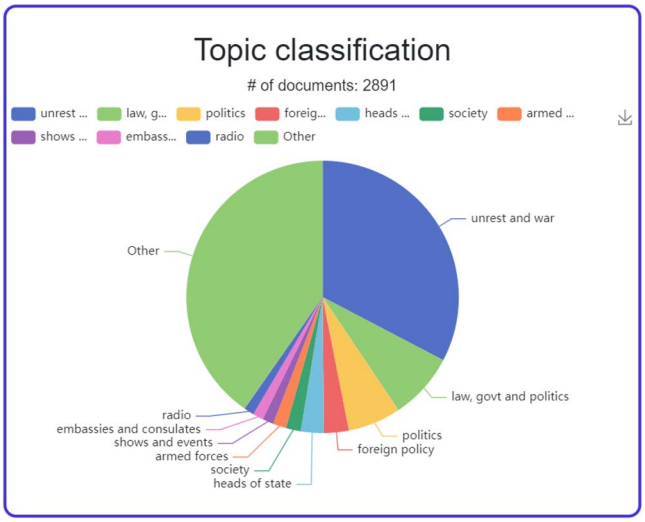

Fig. 15. Russian dataset topics.

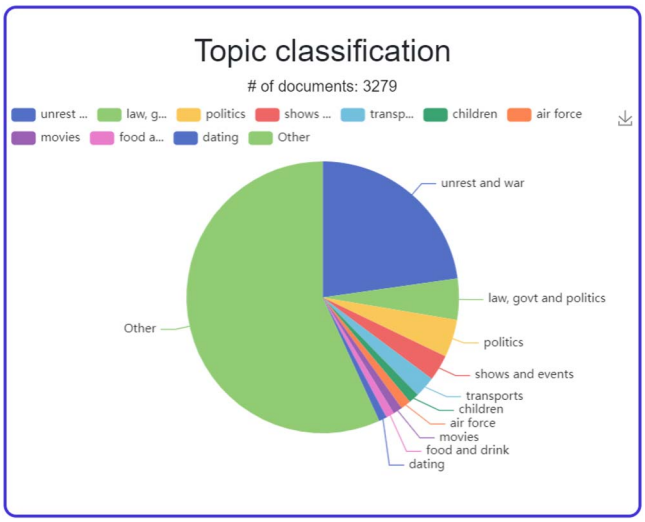

Fig. 16. Ukrainian dataset topics.

VI. Conclusion and Future Work

This article introduces a novel online tool called DataPoll. Its aim is to bridge the gap between two research areas, that is, tools that collect and sometimes analyze isolated data, and isolated research that applies computational methods on specific topics. Furthermore, it aims at providing new techniques that can be applied on big data and open new paths in computational social science research. We strive to achieve that by proposing and implementing an online platform called DataPoll offering some unique and innovative features that no other tool offers by

1) Bringing together all major analyses currently used by social scientists.
2) Facilitating multisource data collection and cross-source data analysis.
3) Facilitating comparative and exploratory visualization to better understand the analysis results.
4) Support community collaboration by allowing end-users to add new features and capabilities.

An example of a use case is also presented, exploring the rhetoric regarding the Russian–Ukrainian conflict. This process involved collecting online content from both Ukrainian and Russian sources, applying various text analysis techniques, and finally utilizing comparative analysis on interactive visualizations to identify differences in the discourse used by the two sides.

Despite DataPoll's effectiveness on the test-case example, there are additional features that must be included, mainly related to new analysis techniques. Currently, most techniques implemented in DataPoll are not domain-specific and have been developed for general natural language tasks that can be applied across fields. Our immediate goal is to develop our own text analysis techniques, tailored exclusively for political content. One field of particular interest is the mining of arguments, which focuses on the identification and detection of the arguments expressed in the text [74]. Knowing the arguments relating to a particular stance of a citizen (e.g., negative sentiment) about a topic (e.g., climate change) can help us better understand why that stance is being held.

In conclusion, DataPoll is not meant to replace traditional methods of analysis, but rather to serve as a tool that: 1) enables the research community to apply big data analysis on text corpora in a simple, uniform, and intuitive manner; 2) supports multidimensional analysis leading to better understanding of the data; and 3) enables addition of new domain specific analysis techniques without having to worry about any other aspects of the big data pipeline.

References

[1] C. H. d. Vreese, "Context, elites, media and public opinion in referendums: When campaigns really matter," in *The Dynamics Referendum Campaigns*. London, UK: Springer, 2007, pp. 1–20.
[2] B. O'Connor, D. Bamman, and N. Smith, "Computational text analysis for social science: Model assumptions and complexity," in *Proc. 2nd Workshop Comput. Social Sci. Wisdom Crowds (NIPS)*, 2011.
[3] J. Wilkerson and A. Casas, "Large-scale computerized text analysis in political science: Opportunities and challenges," *Annu. Rev. Political Sci.*, vol. 20, no. 1, pp. 529–544, 2017.

[4] K. Benoit, *The SAGE Handbook of Research Methods in Political Science and International Relations* (Text as Data: An Overview), vol. 2. London, U.K.: SAGE, 2020, pp. 461–497.

[5] R. M. Chang, R. J. Kauffman, and Y. Kwon, "Understanding the paradigm shift to computational social science in the presence of big data," *Decis. Support Syst.*, vol. 63, pp. 67–80, Jul. 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923613002212

[6] R. Conte et al., "Manifesto of computational social science," *Eur. Phys. J. Special Topics*, vol. 214, no. 1, pp. 325–346, 2012.

[7] W. van Atteveldt and T.-Q. Peng, "When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science," *Commun. Methods Measures*, vol. 12, nos. 2–3, pp. 81–92, 2018.

[8] Y. Goldberg, *Neural Network Methods for Natural Language Processing* (*Synthesis Lectures on Human Language Technologies*), 2017, vol. 10, no. 1, pp. 1–309.

[9] S. Stephens-Davidowitz, *Everybody Lies: What the Internet Can Tell Us About Who We Really Are*. London, UK: Bloomsbury Publishing, 2018.

[10] V. Kagan, A. Stevens, and V. Subrahmanian, "Using Twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election," *IEEE Intell. Syst.*, vol. 30, no. 1, pp. 2–5, Jan./Feb. 2015.

[11] A. Karami, L. S. Bennett, and X. He, "Mining public opinion about economic issues: Twitter and the US presidential election," *Int. J. Strategic Decis. Sci.*, vol. 9, no. 1, pp. 18–28, 2018.

[12] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Syst. Appl.*, vol. 116, pp. 209–226, Feb. 2019.

[13] M. Chopra, S. K. Singh, A. Gupta, K. Aggarwal, B. B. Gupta, and F. Colace, "Analysis & prognosis of sustainable development goals using big data-based approach during COVID-19 pandemic," *Sustain. Technol. Entrepreneurship*, vol. 1, no. 2, 2022, Art. no. 100012.

[14] G. G. Devarajan, S. M. Nagarajan, S. I. Amanullah, S. S. A. Mary, and A. K. Bashir, "AI-assisted deep NLP-based approach for prediction of fake news from social media users," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4975–4985, Aug. 2024.

[15] A. Amira, A. Derhab, S. Hadjar, M. Merazka, M. G. R. Alam, and M. M. Hassan, "Detection and analysis of fake news users' communities in social media," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 5050–5059, Aug. 2024.

[16] C. Cioffi-Revilla, "Introduction to computational social science," Heidelberg, Germany: Springer-Verlag, 2014.

[17] D. Marciniak, "Computational text analysis: Thoughts on the contingencies of an evolving method," *Big Data Soc.*, vol. 3, no. 2, 2016, Art. no. 2053951716670190.

[18] W. Lowe, K. Benoit, S. Mikhaylov, and M. Laver, "Scaling policy preferences from coded political texts," *Legislative Stud. Quart.*, vol. 36, no. 1, pp. 123–155, 2011.

[19] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song, "The politics of comments: Predicting political orientation of news stories with commenters' sentiment patterns," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2011, pp. 113–122.

[20] D. Greene and J. P. Cross, "Exploring the political agenda of the European parliament using a dynamic topic modeling approach," *Political Anal.*, vol. 25, no. 1, pp. 77–94, 2017.

[21] H. Roberts et al., "Media cloud: Massive open source collection of global news open web," in *Proc. Int. AAAI Conf. Web Social Media (ICWSM)*, 2021, pp. 1034–1045.

[22] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: Learning about world events from news," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 107–110.

[23] T. Willaert, P. Van Eecke, K. Beuls, and L. Steels, "Building social media observatories for monitoring online opinion dynamics," *Social Media Soc.*, vol. 6, no. 2, 2020, Art. 2056305119898778.

[24] Y. Theocharis and A. Jungherr, "Computational social science and the study of political communication," *Political Commun.*, vol. 38, nos. 1–2, pp. 1–22, 2021.

[25] R. Stewart, "Big data and Belmont: On the ethics and research implications of consumer-based datasets," *Big Data Soc.*, vol. 8, no. 2, 2021, Art. no. 20539517211048183.

[26] D. Hillard, S. Purpura, and J. Wilkerson, "Computer-assisted topic classification for mixed-methods social science research," *J. Inf. Technol. Politics*, vol. 4, no. 4, pp. 31–46, 2008.

[27] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Anal.*, vol. 21, no. 3, pp. 267–297, 2013.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, pp. 993–1022, Mar. 2003.

[29] D. M. Blei and J. D. Lafferty, "Topic models," in *Text Mining*. London, U.K.: Chapman and Hall, 2009, pp. 101–124.

[30] M. E. Roberts et al., "The structural topic model and applied social science," in *Proc. Adv. Neural Inf. Process. Syst. Workshop Topic Models: Comput., Appl., Eval.*, vol. 4. Harrahs and Harveys, Lake Tahoe, 2013, pp. 1–20.

[31] B. Liu, *Sentiment Analysis and Opinion Mining* (*Synthesis Lectures on Human Language Technologies*), 2012, vol. 5, no. 1, pp. 1–167.

[32] D. Kang and Y. Park, "Based measurement of customer satisfaction in mobile service: Sentiment analysis and vikor approach," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1041–1050, 2014.

[33] Y.-M. Li and T.-Y. Li, "Deriving market intelligence from microblogs," *Decis. Support Syst.*, vol. 55, no. 1, pp. 206–217, 2013.

[34] H. Rui, Y. Liu, and A. Whinston, "Whose and what chatter matters? The effect of tweets on movie sales," *Decis. Support Syst.*, vol. 55, no. 4, pp. 863–870, 2013.

[35] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[36] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," 2016, *arXiv:1603.01360*.

[37] D. Mollá, M. Van Zaanen, and D. Smith, "Named entity recognition for question answering," in *Proc. Australas. Lang. Technol. Workshop*, 2006, pp. 51–58.

[38] M.-G. J. Okurowski, C. Aone, and I. Larsen, "A trainable summarizer with knowledge acquired from robust NLP techniques," in *Advances in Automatic Text Summarization*, I. Mani, and M. T. Maybury, Eds., MA, USA: MIT Press, 1999, pp. 4–5.

[39] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022.

[40] M. Laver, K. Benoit, and J. Garry, "Extracting policy positions from political texts using words as data," *Amer. Political Sci. Rev.*, vol. 97, no. 2, pp. 311–331, 2003.

[41] J. B. Slapin and S.-O. Proksch, "A scaling model for estimating time-series party positions from texts," *Amer. J. Political Sci.*, vol. 52, no. 3, pp. 705–722, 2008.

[42] J. Scott and P. J. Carrington, *The SAGE Handbook of Social Network Analysis*. CA, USA: SAGE, 2011.

[43] J. Scott, "Social network analysis: Developments, advances, and prospects," *Social Netw. Anal. Mining*, vol. 1, no. 1, pp. 21–26, 2011.

[44] D. Camacho, Á. Panizo-LLedot, G. B. Orgaz, A. González-Pardo, and E. Cambria, "The four dimensions of social network analysis: An overview of research methods, applications, and software tools," 2020, *arXiv:2002.09485*.

[45] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust; IEEE 3rd Int. Conf. Social Comput.*, Piscataway, NJ, USA: IEEE Press, 2011, pp. 192–199.

[46] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci.*, vol. 99, no. 12, pp. 7821–7826, 2002.

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Statist. Mech. Theory Exp.*, vol. 2008, no. 10, 2008, Art. no. P10008. [Online]. Available: http://stacks.iop.org/1742-5468/2008/i=10/a=P10008

[48] A. Charalampous, C. Djouvas, N. Tsapatsoulis, and E. Kouzaridi, "Emerging research topics identification using temporal graph neural networks," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innovations*, London, UK: Springer, 2024, pp. 192–205.

[49] H. Partaourides, E. Kouzaridi, N. Tsapatsoulis, and C. Djouvas, "On the identification of influential topics in the social sciences using citation analysis," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput./Int. Conf. Pervasive Intell. Comput./Int. Conf. Cloud Big Data Comput./Int. Conf Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 0845–0851.

[50] N. Iliinsky and J. Steele, *Designing Data Visualizations: Representing Informational Relationships*. CA, USA: O'Reilly Media, Inc., 2011.
[51] D. Lazer, D. Brewer, N. Christakis, J. Fowler, and G. King, "Life in the network: The coming age of computational social," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.
[52] S. S. Roy, A. Roy, P. Samui, M. Gandomi, and A. H. Gandomi, "Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 5028–5037, Aug. 2024.
[53] S.-O. Proksch, W. Lowe, J. Wäckerle, and S. Soroka, "Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches," *Legislative Stud. Quart.*, vol. 44, no. 1, pp. 97–131, 2019.
[54] C. Cochrane, L. Rheault, J.-F. Godbout, T. Whyte, M. W.-C. Wong, and S. Borwein, "The automatic analysis of emotion in political speech based on transcripts," *Political Commun.*, vol. 39, no. 1, pp. 98–121, 2022.
[55] L. Hemphill and A. M. Schöpke-Gonzalez, "Two computational models for analyzing political attention in social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, 2020, pp. 260–271.
[56] V. Triga, F. Mendez, and C. Djouvas, "Post-crisis political normalisation? The 2018 presidential elections in the Republic of Cyprus," *South Eur. Soc. Politics*, vol. 24, no. 1, pp. 103–127, 2019.
[57] V. Triga, N. Ioannidis, and C. Djouvas, "The waning of ideology? Presidential elections in the Republic of Cyprus, 5 February 2023," *South Eur. Soc. Politics*, vol. 28, no. 2, pp. 177–206, 2023.
[58] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2187–2195, 2023.
[59] D. Walter and Y. Ophir, "Strategy framing in news coverage and electoral success: An analysis of topic model networks approach," *Political Commun.*, vol. 38, no. 6, pp. 707–730, 2021.
[60] N. Kligler-Vilenchik, M. de Vries Kedem, D. Maier, and D. Stoltenberg, "Mobilization vs. demobilization discourses on social media," *Political Commun.*, vol. 38, no. 5, pp. 561–580, 2021, doi: 10.1080/10584609.2020.1820648.
[61] L. Hu and M. W. Kearney, "Gendered tweets: Computational text analysis of gender differences in political discussion on Twitter," *J. Lang. Social Psychol.*, vol. 40, no. 4, pp. 482–503, 2021.
[62] M. Zehring and E. Domahidi, "German corona protest mobilizers on Telegram and their relations to the far right: A network and topic analysis," *Social Media+ Soc.*, vol. 9, no. 1, 2023, Art. no. 20563051231155106.
[63] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?" *Comput. Linguistics*, vol. 50, no. 1, pp. 237–291, 2024.
[64] W. Van Atteveldt, J. Strycharz, D. Trilling, and K. Welbers, "Computational communication science—Toward open computational communication science: A practical road map for reusable data and code," *Int. J. Commun.*, vol. 13, 2019, Art. no. 20.
[65] K. Leetaru and P. A. Schrodt, "GDELT: Global data on events, location, and tone, 1979–2012," in *ISA Annu. Conv.*, vol. 2, no. 4. Citeseer, 2013, pp. 1–49.
[66] D. Deacon, "Yesterday's papers and today's technology: Digital newspaper archives and 'push button' content analysis," *Eur. J. Commun.*, vol. 22, no. 1, pp. 5–25, 2007.
[67] D. Li et al., "ECharts: A declarative framework for rapid construction of web-based visualization," *Vis. Inform.*, vol. 2, no. 2, pp. 136–146, 2018.
[68] D. Merkel, "Docker: Lightweight linux containers for consistent development and deployment," *Linux J.*, vol. 2014, no. 239, 2014, Art. no. 2.
[69] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, "Borg, Omega, and Kubernetes," *Commun. ACM*, vol. 59, no. 5, pp. 50–57, Apr. 2016, doi: 10.1145/2890784.
[70] H. Zeng, Z. Su, Q. Xu, and R. Li, "Security and privacy in space-air-ocean integrated unmanned surface vehicle networks," *IEEE Netw.*, vol. 38, no. 3, pp. 48–56, May 2024.
[71] N. Kligler-Vilenchik, C. Baden, and M. Yarchi, "Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and Whatsapp," *Social Media+ Soc.*, vol. 6, no. 3, 2020, Art. no. 2056305120944393.
[72] K.-C. Yang et al., "The COVID-19 infodemic: Twitter versus Facebook," *Big Data Soc.*, vol. 8, no. 1, 2021, Art. no. 20539517211013861.
[73] M. J. Denny and A. Spirling, "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it," *Political Anal.*, vol. 26, no. 2, pp. 168–189, 2018.
[74] J. Lawrence and C. Reed, "Argument mining: A survey," *Comput. Linguistics*, vol. 45, no. 4, pp. 765–818, 2020.

Antonis Charalampous is currently working toward the Ph.D. degree in computational social systems with the Department of Communication and Internet Studies, Cyprus University of Technology, Limassol, Cyprus.

His research interests include digital methods, big data mining, and natural language understanding, with a special focus on developing new techniques and systems for analyzing online discourse.

Constantinos Djouvas received the Ph.D. degree in computer science from The City University of New York, New York, NY, USA.

Currently, he is an Assistant Professor with the Communication and Internet Studies Department, Cyprus University of Technology, Limassol, Cyprus. His research interests include designing and developing innovative computational techniques for collecting, processing, and analyzing large-scale heterogeneous data to gain insights into sociocultural phenomena, with an emphasis on data visualization. He has contributed to several EU-funded projects, including the H2020 COHESIFY study on media representations of Cohesion Policy and the H2020 RePast project, where he built a multilevel interactive data platform.

Christos Christodoulou received the master's degree in data science and engineering in 2023, from Cyprus University of Technology, Limassol, Cyprus, where he is currently working toward the Ph.D. degree in computational social systems with the Department of Communication and Internet Studies.

His research interests include machine learning and deep learning, particularly graph neural networks for recommender systems. He has participated in EU-funded research projects. With industry experience in leading AI-driven software products, he is passionate about applying advanced machine learning to solve real-world challenges.