Berkovitz *et al. Journal of Big Data* (2023) 10:161
https://doi.org/10.1186/s40537-023-00831-3

Journal of Big Data

RESEARCH Open Access

Open framework for analyzing public parliaments data

Shai Berkovitz[1*†], Amit Mazuz[1†] and Michael Fire[1]

†Shai Berkovitz and Amit Mazuz have contributed equally to this work.

*Correspondence: shaiber@post.bgu.ac.il

[1] Ben-Gurion University, Beer-Sheva, Israel

Abstract

Open information about government organizations should interest all citizens who care about their governments' functionality. Large-scale open governmental data open new opportunities for citizens and researchers to monitor their government's activities and improve its transparency. Over the years, various projects and systems have processed and analyzed governmental data based on open government information. Here, we present the Collecting and Analyzing Parliament Data (CAPD) framework. This novel generic open framework enables collecting and analyzing large-scale public governmental data from multiple sources. This study utilized our framework to collect over 64,000 parliament protocols from over 90 committees from three countries and analyzed it to calculate structured features. Next, we utilized anomaly detection and time series analysis to achieve a number of insights into the committees' activities. This study demonstrates that the CAPD framework can be utilized to effectively identify anomalous meetings and detect dates of events that affect the parliaments' functionality and help to monitor their activities.

Keywords: Open government, Data science, Parliamentary monitoring, Big data

Introduction

Open information about government organizations is of interest to all citizens who care about their governments' functionality. According to Hulstijn et al. [1], there are five key issues that are essential for open information: government transparency, improving public service, innovation, economic value, and efficiency. Moreover, open information is crucial to various fields, such as health, emergency, and transportation.

Additionally, public data is a useful source for of social innovation and economic growth. Open data provides new opportunities for governments to collaborate, and for businesses and entrepreneurs to understand potential markets better and build data-driven products [2].

Over the years, various projects and systems have processed and analyzed governmental data using open government information. For example, the Openkamer *Openkamer*[1]

[1] https://github.com/openkamer/openkamer.

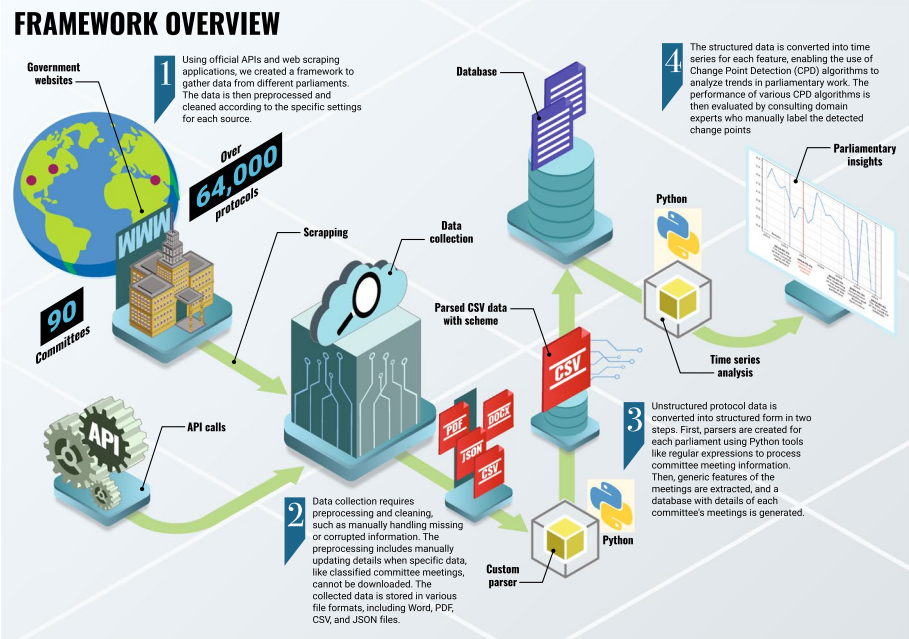


Fig. 1 Framework overview

project analyzes public datasets in the Netherlands to provide insights into the procedures in the Dutch parliament. Other projects focus on collecting and processing government data by public APIs or web scrapping and making it more accessible to the public, such as the *Open Knesset*[2] project, which mines the Israel parliament (Knesset) activities from the official Knesset website to track voting, legislation, and committee activities.

This study presents the Collecting and Analyzing Parliament Data framework (CAPD) framework. This generic novel framework offers data analysis of govern - mental and parliamentary information presented as time series. This allows us to gain insights identify anomalies regarding the organization's work habits and functionality during certain events or periods. Our framework is built from three primary components (see Fig. 1): the first component collects parliamentary data from different countries using public APIs and custom web scrappers. The second aspect analyzes the collected data and extracts features from it. Lastly, it performs a time series analysis of the extracted features and to generate insights into the parliament's functionality within a specific period. We evaluated our framework's ability to identify events and changes in parliament committees' activities by collecting large-scale parliamentary data from the USA, Canada, and Israel. Moreover, using the framework, we analyzed the curated data using machine learning algorithms (see Sect. "Methods and experiments"). The results show that the open CAPD framework can collect and analyze a large amount of data and detect dates of events that affect parliaments' functionality, such as the COVID-19 pandemic and special discussions in specific committees (see Sect. "Results").

[2] https://oknesset.org.

This paper presents three key contributions:

- We introduce a novel open generic framework that can be used to collect and analyze large-scale parliamentary data from different parliaments worldwide. The framework can be utilized to collect and analyze additional parliamentary data and advance data-driven research of governments' activities and can be used to monitor parliament activities.
- We curated a large-scale open data set with over 64,000 protocols collected from different countries.
- We demonstrate that using Change Point Detection algorithms (see Sect. "Time series analysis") can be used to analyze parliamentary data and detect significant committee meetings with relatively high precision (see Sect. "Results").

The remainder of this article is organized as follows. Section "Related work" offers an overview of open government information publicity and its contribution to government activity. Additionally, we examine the related systems that analyze government information and give some knowledge about the time series analysis used in our research. Section "Methods and experiments", describes our methodology for building the CAPD framework and our evaluation of the various methods. Section "Results" details the different results of our collected data and analyzing methods to compare the analysis algorithms. In Sect. "Discussion", we discuss the obtained results. Lastly, Sect. "Conclusions", presents the study's conclusions and future research directions.

Related work

This Section provides an overview of related works to our multidisciplinary study. First, in Section "Open government information", we review and detail the disclosure of information in different governments and parliaments worldwide. Next, Sect. "Contribution to government activity", describes the contributions open information in governments can make. Section "Systems for analyzing government information", reviews the various parliament systems and what they can offer the public, and Sect. "Data mining parliamentary data", provides a short overview of studies in which data mining algorithms analyze open government data (OGD). Finally, Sect. "Time series analysis", presents several methods for time series analysis used in this study.

Open government information

As it has developed in the Western world in recent years, Open government policy, seeks to harness new technologies. It does so for the benefit of improving the means of communication between government and its citizens and for the proper utilization of the social and economic benefits inherent in the information held by government authorities [3]. Public sector information technologies are used at the federal and regional levels to improve the efficiency of government bodies' administrative activities [3].

One of the main benefits of promoting transparency, access to information, and public participation in the digital age is the ability to monitor the government's actions [4]. Another advantage is that access to information in various fields makes it possible to derive public and economic benefits and encourage development and entrepreneurship.

Of course, public participation is an excellent tool for implementing this concept and a principle of good governance. It encourages accountability and supports wider participation, inviting contributions from the public as well as experts [4].

According, to Barbra Ubaldi [5], governments must seek feedback from the public on the usefulness, relevance, and accessibility of their data to enable continuous improvement. Furthermore, many civil servants see the real-time performance and impact of public services and public policy on citizens. They can create appropriate data and other inputs, or use available data, to improve the service experience, This is dependent on the provision of tools and incentives, such as the opportunity to participate in online social networks in a professional role in order to offer advice and information to the public [5]. Therefore, governments need to recognize the value of the audience source to enable real-time data and information sharing and engage relevant stakeholders outside public organizations to use them to create value [5].

Many governmental datasets are accessible to the public by several interfaces: (a) *requesting information* on relevant government websites; (b) consuming the information Application Programming Interfaces (APIs)[3] provided by the government to the public; or (c) by *crawling and scraping* various government websites. There are some several examples:

Applying online - FOIA FOIA is a UK website[4] that details how to apply: the request for information can be submitted by sending a written request to any public body, who will reply within 20 days. Most applications are free of charge, however some may require payments to submit the request. In the US and the UK, the body (agency) to which the application should be sent (each body handles the request) must be identified. It is important to note that the information seeker pays for search hours. The public body commits to deadlines according to the body and information requested. If the request is too extensive, the body can reject it.

Using API access In the UK, the API approach is currently in the ALPHA stages.[5] However, it offers about 200 access points (URL endpoints) on 26 topics. Each access point contains documentation and an API. In the US, access to information through the API is quite advanced and high-quality. The API offers various data from 40 access points on 18 topics.

Crawling and scrapping online data Data scraping is the process of extracting data from a human-readable output from another program. Web crawling is a type of data scraping that aims to explore the web, primarily finding URLs and links. Web scraping is also a type of data scraping that extracts data and information from web-pages [6]. In the USA, you can crawl for links to all the committees' pages and can scrape valuable data like congressional bills, reports, hearings, and prints.[6] In Israel, you can scrape all the Knesset protocols by writing a scrapper to a single web-page.[7]

[3] Application Programming Interface (API) is a set of code libraries or pre-made functions, which programmers can make easy use of, without having to write them themselves so that they can use the information of the application they want to use for their application.

[4] https://www.foia.gov/.

[5] https://www.api.gov.uk/.

[6] https://www.govinfo.gov/browse/committee.

[7] https://main.knesset.gov.il/Activity/committees/Pages/AllCommitteeProtocols.aspx.

Recently, several federal data monitoring and conservation projects have been conducted that are scientifically hosted on government databases in the US and other countries [7].

Today, governments are trying to be more "open."[8] One aspect of open government is opening governmental data [8–10]. However, it is known that simply providing open government data does not automatically produce significant societal value [8]. Research in this area often cites the many potential benefits of OGD [8, 11–13]. However, these benefits will not be realized unless the data is used. Thus, a concrete understanding of the barriers that prevent OGD use to produce public value is essential. As a continuation of this, a framework is needed that will guide the use of OGD efficiently and effectively to that produces as much public value as possible [11].

Contribution to government activity

Governments collect data for future use. For example, during the COVID-19 pandemic, researchers and government employees used existing information to support their initiatives and responses [14]. The demand for, relevant, and quality data access has increased. This has been driven by taking informed and rapid policy action, improving engagements, conducting a scientific analysis of a dynamic threat, and understanding the social and economic impacts that enable oversight and reporting on Civil Society. Data and AI are innovative components that can help find solutions to social challenges, from health to agriculture and security to manufacturing [14]. On the other hand, the multiple uses of data must be balanced to responsibly maintain high standards of privacy, security, safety, and ethics [15].

For example, the World Health Organization (WHO) stated that they used open data and location analytics to eliminate malaria in their annual report.[9] Thousands of volunteers have worked on mapping the Malaria-affected world, leading to a more efficient distribution of resources and identification of the areas in need of the most support. The collection and use of improved data contributed to an 85% decrease in reported Malaria cases and a 92% decrease in Malaria-related deaths across the southern province, affecting approximately 1.8 million people.[??]

Researchers can use government data to draw groundbreaking conclusions, as demonstrated by the treatment of the COVID-19 pandemic [16]. Although large amounts of government data have been made available through databases and portals, there needs to be more evidence of dedicated services or innovations created from OGD reuse. In line with broader calls to ensure a "targeted advertising" approach, organizations need to understand better the needs of data across the ecosystem [14].

In 2015, it was expected that using OGD would save European governments 1.7 billion euros and produce 350,000 jobs by 2020 [7]. These goals have been achieved, as seen in a report published in the European Data Portal in 2020 [17]. One of the reasons for achieving these goals is the number of resources Europe invests in open-data, particularly in OGD. In fact, all but one of the parliaments in Europe have a body whose primary purpose is to promote the use and quality of OGD [15]. These bodies coordinate

[8] https://www.ipc.nsw.gov.au/media/3241.

[9] https://apps.who.int/iris/bitstream/handle/10665/275867/9789241565653-eng.pdf?ua=1.

open data initiatives and create supporting materials to publish and reuse available data, and also provide training to civil servants.

Systems for analyzing government information

More bodies, public institutions, and governments have acted accordingly and adapting to the use of available data in recent years. Today, numerous data sets are offered to the general public in many countries, including more relevant and usable information.

For example, the CA and the US parliaments' portals contain over 25,000[10] and 330,000[11] data-sets, respectively.

Additionally, there have been different open code initiatives utilizing available governmental data sets. The *Openkamer* code project[12] provides insight into the Dutch parliament by extracting parliamentary data from several external sources. It visualizes this data in a web application, such as legislative proposals, queries, political parties, and gifts to parliament. Similarly, in the UK, the Public Whip project,[13] ais an independent non-governmental project that, web-scrapes the House of Commons and House of Lords debate transcripts to enable the public to monitor and influence voting patterns.[14] Similar projects have been implemented in France and Canada. In France, the *senapy* Python project[15] scrapes data from the French Senate website,[16] and in Canada, the *coalition-analyzer*[17] code project submits open API requests for the Canadian House of Representatives, so as to calculate correlations in the votes of different parties. In Israel, the *Open Knesset*[18] project mines all Israel parliamentary activity from the official Knesset website to track voting, legislation, and committee activities.

Data mining parliamentary data

Several studies have taken public parliamentary information and applied natural language processing (NLP) methods to detect political ideology, sentiment, position-taking, perspectives, and the level of agreement and disagreement between politicians.

In 2003, Laver and Bendit [18] presented a method for extracting policy positions from political text. They start by implementing their method and testing it on Britain and Ireland's political parties, before "exporting" their model to non-English languages like German, extending the model from party manifesto analysis to estimate political positions from legislative speeches.

Awadallah et al. [19] presented *OpinioNetIt*, in 2021, which is a structured, faceted knowledge-base of opinions, which tracks its use in political views analysis.

[10] https://open.canada.ca/en/open-data.

[11] https://data.gov.

[12] https://github.com/openkamer/openkamer.

[13] https://www.publicwhip.org.uk.

[14] https://www.publicwhip.org.uk/faq.php.

[15] https://github.com/regardscitoyens/senapy.

[16] https://www.senat.fr/.

[17] https://github.com/oliversno/coalition-analyzer.

[18] https://oknesset.org.

They focused on acquiring opinions held by various stakeholders on politically controversial topics. Awadallah et al.'s system can be used for multiple types of analysis, including heatmaps showing political bias, flip-flopping politicians, and dissenters [19].

In 2014, Iyyer et al. [20] implemented a Recurrent Neural Network (RNN) for identifying the political position indicated by a single sentence. They used multiple public data sets containing over a million sentences. They then filtered, annotated, and processed them to adjust their identification task, resulting in almost 100,000 labeled sentences.

In 2017, Vilares proposed "a Bayesian modeling approach where topics (or propositions) and their associated perspectives (or viewpoints) are modeled as latent variables" [21]. They evaluated their model on debates from UK's House of Commons scrapped using a custom web crawler, revealing perspectives from debates without labeled data. In the same year, Gencheva et al. [22] constructed an entire corpus of political debates containing statements that reputable sources fact-check. Then, they trained machine learning models to predict which claims should be prioritized for fact-checking. Abercrombie and Batista-Navarro [23] published a literature review of 61 studies which addressed the automatic analysis of sentiment, opinions and positions taken by speakers in parliamentary (and other legislative) debates."

In 2020, Stavropoulou et al. [24] introduced the ManyLaws platform, an innovative Legal Web platform facilitating easy access to open legal data in the EU. The platform offers user-centric services, leveraging data analytics and semantic analysis for multilingual resources. Co-designed with legal stakeholders, ManyLaws enhances transparency and insights into legislative activities, exemplifying the potential of open government data analysis. In the same year, Cantador and Sánchez [25] presented a new approach to semantic annotation and retrieval of parliamentary content using ontology-based methods. They build a knowledge base from the Parlamento 2030 dataset[19] and develop a semantic annotation framework. By incorporating a semantic relationship between ontology entities, their retrieval method improves document ranking. The experiments show that their semantic approaches outperform traditional methods. Recently, Varlamis et al. [26] explored the application of recommender systems in parliamentary contexts, focusing on the Hellenic Parliament as a case study. They proposed a prototype design for a pilot search engine powered by a recommender system that recommends relevant parliamentary content to users.

In addition to NLP for analyzing parliamentary data, other studies utilized network science for data mining parliamentary data.

In 2005, Porter et al. [27] investigated the US House of Representatives network of committees and subcommittees. They showed that network theory, combined with the analysis of roll-call votes using singular value decomposition, successfully uncovers political and organizational correlations between committees in the House without the need to incorporate other political information.

Dal Maso et al. [28] conducted a network analysis study that analyzed the network of relations between parliament members according to their voting behavior. As a case study, Dal Maso et al. focused on the Chamber of Deputies of the Italian Parliament.

[19] https://www.parlamento2030.es/about-en.

They found sharp contrasts within the political debate, which does not imply a relevant structure based on established parties. In addition, they introduced a way to track the stability of the government coalition over time, which can discern the contributions of each member and the impact of possible defections. In 2020, an open-source tool was developed to inspect and explore the Israeli parliament.[20] The tool allows the ability to search for various topics by keywords, see parliament members' votes and activities. This tool enables citizens to explore all the popular topics discussed in parliament sessions, and see politicians' opinions and interests.

Lastly, a data set named LOCALVIEW [29] was published in 2023. With its unprecedented scale and coverage, LOCALVIEW is a valuable tool for studying American local government policy-making. This data set consists of over 139,000 videos, as well as corresponding textual and audio transcripts of local government meetings publicly uploaded to YouTube. It covers 1,012 places and 2,861 distinct governments across the US from 2006 to 2022. LOCALVIEW's ability to provide real-time access to regional policy-making, which has been difficult and expensive to study at scale, is remarkable. This data set can aid scholars, journalists, and observers of regional politics and policies in exploring substantive phenomena of their interest. It covers a wide range of municipalities and counties, where its applicability extends beyond local policy-making. It has implications for the study of deliberative democracy, interpersonal communication, and intergroup dynamics along partisan, racial, geographical, or other dimensions.

Time series analysis

This study analyzed data from time series observations at a given time interval, using offline Change Point Detection (CPD) algorithms. CPD presents the problem of estimating the point at which statistical properties of a sequence of observations change. Over the years, several multiple-change-point search algorithms have been proposed to overcome this challenge [30]. CPD is mainly used for two purposes: (a) to identify abnormal sequences along with a time series; and (b) to identify sudden changes in real-time [31].

Detecting those change points is challenging for many applications in different areas, from finance [32] to Bio-informatics [33]. For the first purpose, offline CPD algorithms are used. Offline CPD algorithms assume we have all the necessary data to process. As a result, the algorithms will find all the change points over time and not just the latest changes [31].

Our study utilized four popular offline CPD algorithms:

- *Pruned Exact Linear Time (PELT)*- The algorithm works by minimizing a cost function over possible numbers and locations of changepoints. It involves a pruning step within the dynamic programming approach, which reduces the computational cost while ensuring an exact minimization of the cost function. The PELT method has a linear computational cost in the number of observations, making it more efficient compared to other methods with quadratic or cubic costs. It can be applied to sta-

[20] https://github.com/SgtTepper/BetaKnessetWeb.

tistical criteria such as penalized likelihood, quasi-likelihood, and cumulative sum of squares [34].

- *The binary segmentation method (BinSeg)-* A technique used for change-point detection and signal segmentation. It starts with an initial detection of a change point in the input signal and then splits it into two sub-segments. The algorithm repeats this process on each sub-segment, if necessary. It can also detect multiple change points by recursively splitting the sub-segments. This low-complex technique works with known and unknown regimes and extends other change-point detection methods [35].
- *Window-sliding (WinSlid)-* The algorithm involves sliding a window over the data and calculating the cost function within each window. By iteratively sliding the window and updating the cost function, the algorithm identifies the optimal number and location of changepoints. The computational cost of the WinSlid algorithm is linear in the number of observations, which makes it more efficient compared to existing methods with quadratic or cubic computational costs [31].
- *Dynamic programming search method (DYNP)* - The technique is used to efficiently solve sub-problems for the fixed number of change points to detect in the time series. The algorithm recursively solves sub-problems by relying on the additive nature of the objective function. The algorithm locates the (exact) minimum of the sum of costs by computing the cost of all sub-sequences of a given signal. It is called "dynamic programming" because the search over all possible segmentation is ordered using a dynamic programming approach [31].

Methods and experiments

Our study concludes the parliamentary function by analyzing and processing the meetings of the parliamentary committees. We analyzed thousands of open protocol documents to understand how parliament works over time. In addition, another goal of the study is to develop a generic platform that allows for easy monitoring of various parliamentary bodies and how they change over time.

Namely, given a data set with n text documents (protocols), in any language, containing information centralizing a parliamentary body's work from k different parliamentary committees over t years. We developed a code framework that analyzes each document and creates a database. We utilize this database to discover and track how the parliamentary body has functioned over the years.

The following subsections provide an overview of developing a generic framework. Our method consists of the following steps (see Figure 1). First, we collected parliamentary data from publicly available sources (see Sect. "Data collection and preprocessing"). Afterward, we parsed the collected data into structured data and extract features, and utilized these features to identify abnormal meetings (see Sect. "Data parsing and feature extraction"). Next, we transformed the structured data into time series per feature to allow us to run CPD algorithms to analyze the data and identify changing trends in the parliamentary work (see Sect. "Data analysis"). Lastly, we evaluated the various CPD algorithms' performances by consulting domain experts to label the detected CPDs manually (see Sect. "Evaluation").

Table 1 General Committee schema

General features	Explanation
Country	Country name
Parliament name	Parliament's name
Parliament ID	Parliament ID
Parliament type	Parliament's type
Committee ID	Committee ID
Committee name	Committee name
Meeting ID	Meeting ID
Title	The meeting's title
Date	The date on which the meeting was held
Committee members	A list of committee members who attended the meeting
Members of Parliament	A list of parliament members who attended the meeting
Document length	The amount of characters in the meeting documentation

Table 2 US Committee schema

General features	Explanation
Serial numbers	The serial numbers of the meeting
Witnesses	A list of witness's names who attended the meeting
Location ID	The location identifier where the meeting occurred

Computational framework for analyzing parliamentary data

Data collection and preprocessing

We developed our framework to collect data from various parliaments using different methods, including official APIs and web scraping applications, which extract the necessary data according to the settings for each source (see Sect. 2.1). After collecting the data, it is preprocessed and cleaned. If the collected data contains corrupted or missing information, the missing data is extracted manually. For example, if some committee meetings are classified, the scrapper will be unable to download the meeting files. Therefore, we will manually update basic features about the meetings, such as committee ID and meeting date. At the end of this step, stage, the collected data will be stored in various formats, such as Word, PDF, CSV, and JSON files.

Data parsing and feature extraction

The next step is to take the unstructured data collected and turn it into structured data. For this purpose, we parsed the data by firstly developing a parser for each parliament that receives information about the parliament committees' meetings as input, collected in various formats. Our parsers primarily utilized public Python libraries to process different types of files. Additionally, tools like regular expressions were used to extract the values according to a shared schema we predefined (see Table 1). Secondly, we constructed a "custom" schema that extends the general schema and enriches it with fields relevant only to the specific parliament (see Table 2) to support the unique properties of the different parliaments.

Table 3 Extracted features

General features	Explanation
Committee ID	The committee ID
Meeting text length	The number of characters in the meeting report
Meeting duration	The meeting duration
Year	The year in which the meeting held
Month	The month in which the meeting held
Number of committee members	The number of committee members who attended the meeting
Number of parliament members	The number of parliament members who attended the meeting

We also extracted and selected generic features that characterize parliamentary meetings from the structured data. The feature selection process was done by examining which features are seen in most meetings. It also examined which features could indicate a particular activity trend by comparing the most common features and differences among parliaments. The features we extracted and selected are described in Table 3.

Data analysis

After parsing the data, we transformed the parsed data to present information about similarities or differences between parliaments or various committees of the same parliament. We analyzed the transformed data to detect trends or anomalies, examining the

- *Number of committee meetings.*
- *Average number of committee members.*
- *Average text length of a meeting.*

For each parliamentary committee (denoted C), such as the US Congress Energy and Commerce Committee, we calculate the distribution of each of the three features described above (denoted f). We use these distributions to identify anomalous protocols based on the interquartile range (IQR). The interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data [36]. It is defined as the difference between the 75^{th} (denoted Q3) and 25^{th} (denoted Q1) percentiles of the data. To detect outliers using this method, we define two values, a lower bound, and an upper bound. These are defined as follows:

$$IQR : Q3 - Q1$$
$$LowerBound : (Q1 - 1.5 \cdot IQR)$$
$$UpperBound : (Q3 + 1.5 \cdot IQR).$$

Any value below the lower bound or higher than the upper bound is considered an outlier. For example, we can calculate the distribution of the number of committee members in the Energy and Commerce Committee's meetings. Then, we can calculate the bounds and identify specific meetings with an odd number of attending committee members.

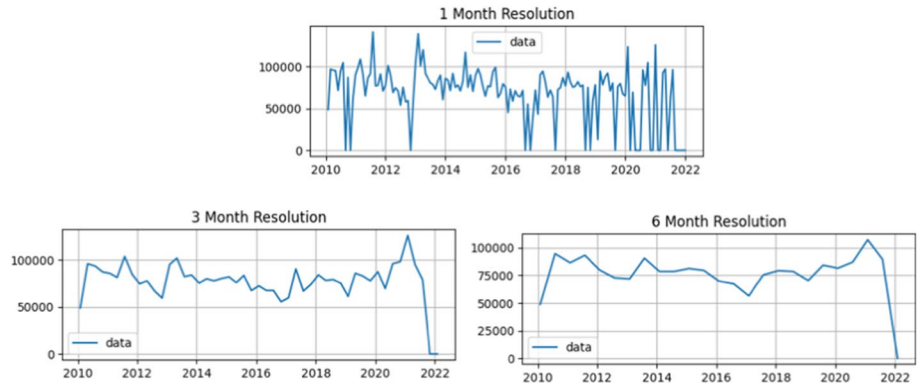


Fig. 2 Israel's Foreign Affairs Committee Protocols' Text Lengths over Time with Different Time Resolutions

For each committee C and each feature f, we generated a time series $g_f^C(t)$, where t is a period. In this study, we chose the values of t to be one of three specific time resolutions: per month, three months, and six months.

As each parliament has a different working schedule and manages its meetings differently, we needed to figure out the proper time resolution to represent our time series.

To overcome this challenge and find the adequate time resolution for representing our data, we plotted graphs for each feature and committee in various time resolutions and charts that compare multiple time series (see Fig. 2).

By observing the variety of generated time series with different time resolutions for each country, we manually selected the resolution that created a suitable time series. We then created three time series for each committee using the appropriate time resolution, one for each feature. On each time series, we run the following CPD algorithms, PELT, BinSeg, Window-based, and DYNP (see Sect. "Time series analysis"). By manually inspecting each algorithm's results for several committees, we will choose two algorithms that perform most adequately.

Evaluation

To evaluate our framework's performance on the different committees, we performed the following evaluations. First, we manually checked the meetings with abnormal values according to the IQR thresholds and calculated true and false positive rates according to the meetings' properties. We evaluated whether the IQR identified anomalous meetings.

Secondly, to evaluate CPD algorithms' performance, we ran the two selected CPD algorithms on the generated time series. Then, for each detected case, we assessed by consulting experts if it was a true or false event. Lastly, we could not evaluate the model using the accuracy or recall metrics because we are unaware of the points (events) we missed detecting, so we assessed each algorithm's performance using *Precision* and *False Positive Rate (FPR)* defined as follow [37]:

$$Precision : \frac{TP}{TP + FP}$$

$$FPR : \frac{FP}{TP + FP}$$

Experimental setup

Data collection

To evaluate our framework and algorithms, we collected parliamentary data from three countries:

- *United States*—In the US, we collected the committees' data of the US House of Representatives by scraping the GovInfo website.[21] GovInfo provides free public access to official publications from all three branches of the Federal Government. GovInfo includes access to many documents, including information on individual publications or collections of content or a view of a list of collections, publications, other resources, and external partner sites. For this study, we collected all available committees' data from January 1999 to May 2021 and parsed them into the framework's schema described in Table 1. In addition, the scrapper can be configured to collect Senate committee data.
- *Canada*—In Canada (CA), we collected data from the House of Commons, the lower house of the Parliament of Canada. We gathered the committees' data by scraping the House of Commons' official website.[22] The House of Commons official website contains diverse data, including Committees' meeting schedules, tentative working agendas, full-length records of what is said in the house, official records of decisions, and more. For this study, we collected all the available committees' data from January 2009 to May 2022 and parsed them into the framework's schema described in Table 1.
- *Israel*—The Israeli parliament's official website[23] provides many data resources regarding the parliament's work. This data includes data about the members of the parliament, transcripts of the plenary and committee sessions, legislative agendas, voting records bills, and more. In this study, we curated a dataset with data collected from the different Knesset committees[24] from January 1999 to January 2021.[25] Next, we analyzed those reports and parsed them into the framework's schema described in Table 1.

After parsing all the data collected for each country, we created time series only for the top 10 most active committees with the highest number of meetings, and that was our curated dataset.

Features distributions

For each country, we looked closely at each country's three most active committees. We calculated the features' distributions for each country's most active committees and the three features. Using the IQR outlier method, we identified meetings with anomalous features (see Sect. "Data analysis"). Then, we manually examined randomly selected

[21] https://www.govinfo.gov/.

[22] https://www.ourcommons.ca/en.

[23] https://main.knesset.gov.il.

[24] https://main.knesset.gov.il/Activity/committees/Pages/AllCommitteeProtocols.aspx.

[25] To make this study reproducible, we also developed a web crawler that can download all these reports.

abnormal meetings in order to determine whether they were genuinely anomalous. Finally, we evaluated this approach's true-positive rates for detecting abnormal features' values.

Change point detection

We focused on the Israeli parliament to evaluate CPD algorithms' performance. As described in Sect. "Data collection and preprocessing", we created time series only for the ten most popular committees, and for each committee, we generated three different time series, one for each feature. Next, we examined our data and found that the suitable time resolution to represent our data was six months of aggregation.

We manually inspected the four CPD algorithm results and found that the PELT and the DYNP algorithms presented adequate results in detecting CPDs. Therefore, we focused on evaluating the PELT and DYNP algorithms.

To accurately evaluate the performance of the PELT and the DYNP algorithms, we needed to check that the change points detected by the algorithms were actual events and not false positive detection. To detect actual events around those dates, we conducted a comprehensive investigation.

We checked and cross-referenced various Knesset systems, applications, databases, protocols, and records. Where possible, we interviewed committee administrators, committee members, and veteran committee directors and asked them what happened at those times. They checked and gave us answers and explanations for cases they knew about.

The classifications range from "no justifiable reason" to "Knesset summer recess/vacation days" and "COVID-19 days and special discussions in specific committees" (see Fig. 4).

Results

As described in Sect. "Data collection", using dedicated web crawlers, we collected committees' data from three countries (see Table 4). We collected the following data: from the US parliament, 16,989 protocols from 21 committees over 22 years; 12,124 protocols from the Canadian parliament from 44 committees over 13 years; and 35,800 protocols from 33 committees for the Israeli parliament over 22 years.

By analyzing the number of collected protocols for each country, we observed that the average and median number of protocols per committee in the US is 809. The average number of protocols per committee in CA is 275, and the median is 416. Lastly, Israel's average number of protocols per committee is 1,084.8, and the median is 226.

Next, as described in Sect. "Data parsing and feature extraction", we analyzed the collected protocols and constructed a detailed database with each committee's meeting details (see Tables 1 and 3). By utilizing the database, for each committee in each country, we calculated the committee's number of meetings, the average number of committee members, and the meetings' protocols' average text lengths (see Table 5 and Fig. 3).

Moreover, we focused on the most active committees. We used the IQR method to uncover anomalous meetings. The IQR algorithm identified 656 anomalous protocols according to three features that were examined. We randomly sampled 263 of these protocols. For the average text length of a meeting feature, we randomly sampled 50

Table 4 Collected parliamentary data overview

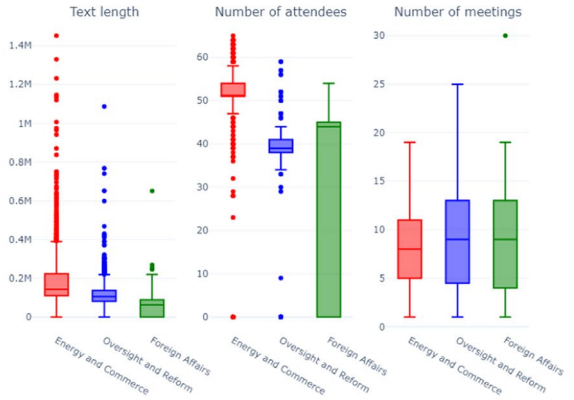
Country	Number of committees	Number of protocols	Time period
US	21	16,989	01/1999–05/2021
CA	44	12,124	01/2009–05/2022
IL	33	35,800	01/1999–01/2021

Table 5 The Committees with the highest number of meetings per country

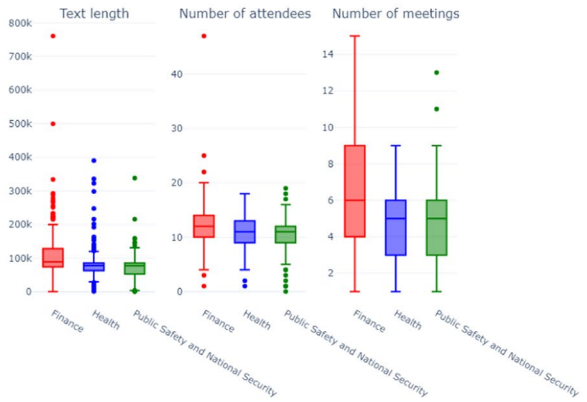
Country	Committee name	Number of meetings
IL	Immigration, absorption and diaspora affairs	1563
	The status of women and gender equality	1302
	Constitution, law and justice	3973
	Economic affairs	4618
	Education, culture and sports	3460
	Finance	5530
	House	2065
	Internal affairs and environment	3501
	Labor and Welfare	3918
	State control	1665
CA	Agriculture and agri-food	481
	Finance	736
	Government operations and estimates	511
	Health	520
	Human resources, skills and social development	507
	Industry, science and technology	487
	Justice and human rights	487
	Procedure and house affairs	508
	Public accounts	474
	Public safety and national security	516
US	Small business	904
	Armed services	1040
	Science, space, and technology	848
	Energy and commerce	1692
	Financial services	1265
	Foreign affairs	1516
	Homeland security and governmental affairs	821
	Oversight and reform	2220
	Natural resources	929
	The judiciary	1130

abnormal protocols from each county. For the average number of committee members feature, we randomly sampled 50 abnormal protocols from the USA and Israel. We also randomly selected six anomalous protocols from Canada.

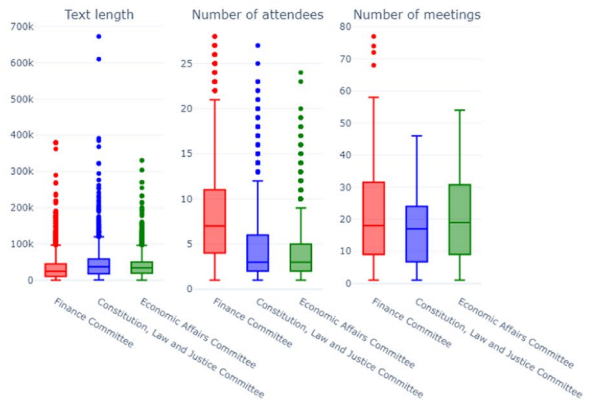
For the number of committee meetings feature (which contained only a few anomalous protocols) we selected one, two, and four protocols for the USA, Canada, and Israel, respectively. Then, we manually examined these sampled protocols and observed that all recognized protocols were indeed abnormal and not a result



(a) US committees feature distributions.



(b) CA committees feature distributions.



(c) IL committees feature distributions.

Fig. 3 Countries' most active committees feature distributions

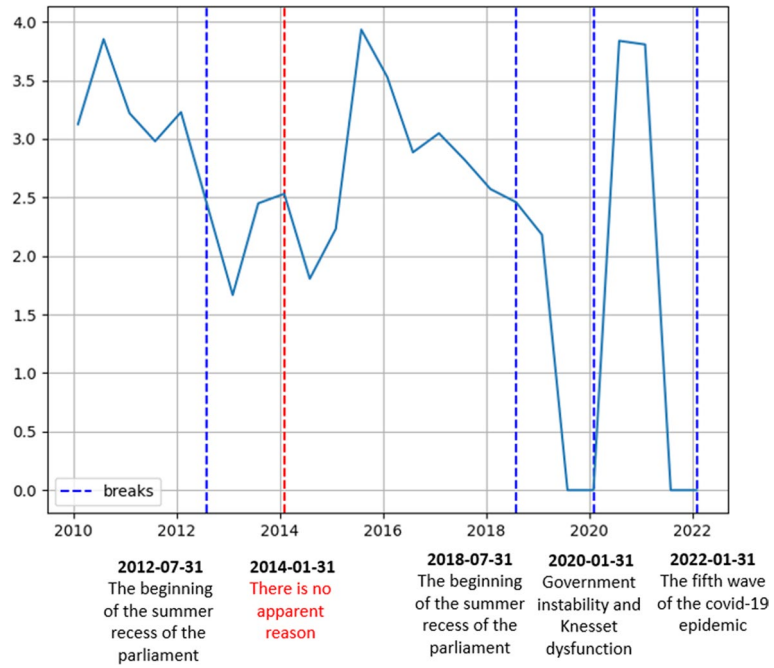


Fig. 4 DYNP detected CPDs on the Knesset's Committee for Immigration, Absorption, and Diaspora Affairs—blue horizontal lines mark true positive events, and the red horizontal line marks a false positive detected event

of problems with the protocols' formats or issues with the parsing of these files.[26] Using this method, we identified interesting and abnormal meetings. For example, we found a charged meeting of the Finance Committee in the Canadian Parliament[27] which discussed the implementation of sections of the budget and was broadcast on TV. In this meeting, the committee chair and members requested the Finance Minister's presence, and not officials on his behalf. Many people participated in this discussion, and the meeting was a relatively long one. Another example is a hearing titled "BP's Role in the Deepwater Horizon Explosion and Oil Spill," which was held by the US House Energy and Commerce Committee, and the Oversight and Investigations subcommittee, which included a long discussion with many people present.

Lastly, we used the PELT and DYNP CPD algorithms and detected events in the Israeli Knesset. We applied the two algorithms with six months of time resolution to all three features.

The two algorithms returned about 289 change points; 124 detected by the PELT algorithm and 165 change points found by DYNP. We then performed a comprehensive investigation (see Sect. "Change point detection") and manually classified each detected point as a true or a false event, and if possible, we added the event's reason. For example, we identified events due to Knesset summer recess, COVID-19

[26] In some cases, we noticed small gaps between the features' values calculated by our algorithms and the features observed in the manual analysis of the protocols. Nevertheless, in most cases, the gaps were relatively minor. For example, we observed slight differences in the protocols' number of characters or minor differences in the number of members' attendees.

[27] https://www.ourcommons.ca/DocumentViewer/en/41-1/FINA/meeting-94/evidence.

Table 6 CPD precision and FPR scores - IL

Committee Name	Data Type	Precision		FPR	
		PELT	DYNP	PELT	DYNP
Immigration, absorption, and diaspora affairs	Number of meetings	0.6	**1**	0.4	**0**
	Number of members	0.5	**0.8**	0.5	**0.2**
	Text length	0.6	0.6	0.4	0.4
Status of women, and gender equality	Number of meetings	0.75	**0.8**	0.25	**0.2**
	Number of members	0.33	**0.8**	0.67	**0.2**
	Text length	0.6	**0.8**	0.4	**0.2**
Constitution, law, and justice	Number of meetings	**0.75**	0.6	**0.25**	0.4
	Number of members	0.67	**0.8**	0.33	**0.2**
	Text length	0.6	0.6	0.4	0.4
Economic affairs	Number of meetings	0.6	**0.8**	0.4	**0.2**
	Number of members	0.5	**0.8**	0.5	**0.2**
	Text length	0.6	0.6	0.4	0.4
Education, culture, and sports	Number of meetings	0.67	**0.8**	0.33	**0.2**
	Number of members	0.5	**0.8**	0.5	**0.2**
	Text length	0.6	**0.8**	0.4	**0.2**
Finance	Number of meetings	0.5	**0.8**	0.5	**0.2**
	Number of members	0.5	**0.8**	0.5	**0.2**
	Text length	0.6	**0.8**	0.4	**0.2**
House	Number of meetings	0.5	**0.8**	0.5	**0.2**
	Number of members	0.5	**0.8**	0.5	**0.2**
	Text length	0.6	0.6	0.4	0.4
Internal affairs, and environment	Number of meetings	**1**	0.6	**0**	0.4
	Number of members	**1**	0.6	**0**	0.4
	Text length	0.6	0.6	0.4	0.4
Laber and welfare	Number of meetings	0.75	**0.8**	0.25	**0.2**
	Number of members	0.5	**0.6**	0.5	**0.4**
	Text length	0.6	**0.8**	0.4	**0.2**
State control	Number of meetings	0.67	**1**	0.33	**0**
	Number of members	0.33	**1**	0.67	**0**
	Text length	0.6	**0.8**	0.4	**0.2**
	Number of meetings	*0.679*	***0.8***	*0.321*	***0.2***
Average score	*Number of members*	*0.533*	***0.78***	*0.467*	***0.22***
	Text length	*0.6*	*0.7*	*0.4*	***0.3***

Bold values indicate the highest values achieved for each metric within each Knesset committee

days, and special discussions in specific committees or government instability and dysfunction of the Knesset (see Fig. 4). Table 6 presents Precision and False Positive values obtained by the PELT and DYNP algorithms across different Knesset committees for each feature. The highest values achieved for each metric within each Knesset committee are highlighted in bold. Furthermore, an overall average for each metric is presented in Table 6 and is marked in italic. In many cased, both algorithms produced high Precision and FPR values. Overall, precision-wise, the DYNP algorithm achieved better results than the PELT algorithm, with an average precision of 0.76 vs. 0.604 (see Table 6).

Discussion

By analyzing the results presented in Sect. "Results", the following can be noted:

Firstly, a vast amount of government data is accessible online. Using different data collection methods described in Sect. "Open government information", we utilized the CAPD framework to collect more than 64,000 government meeting protocols from three countries over more than twenty years (see Table 4). This accumulated unstructured data corpus can help us study various government activities in different geographical locations over extended periods.

Secondly, even though the collected data were from different countries, by using our framework it was feasible to convert the collected unstructured data into structured data (see Sect. "Data parsing and feature extraction" and Tables 1, 2, and 3). This makes it possible to use the proposed framework to gather data from various countries. This makes it possible to compare countries' activities to specific events, such as climate change or emergency.

Thirdly, we demonstrated that it is possible to use data science tools to analyze the collected protocols to monitor governmental committees' activities over time. Making this type of analysis open to the public can help make governmental proceedings more transparent. Moreover, it can also help government members monitor their activities.

Fourthly, it is possible to analyze the structured parliamentary data with anomaly detection algorithms, such as IQR, to identify abnormal meetings, such as anomalous meetings with many participants or extensive duration (see Sect. "Results"). This algorithm can be utilized to detect periods in time with a relatively high number of committee meetings and identify interesting meetings or meetings with public interest, such as the "BP's Role in the Deepwater Horizon Explosion and Oil Spill" hearing (see Sect. "Results").

Fifthly, by employing a developed framework, we showcased the potency of even a small set of calculated features across extended timeframes. Notably, despite the modest number of features, our investigation into their evolution over the years yielded diverse insights. Notably, our approach successfully pinpointed anomalous meetings (see Fig. 3, Sects. "Computational framework for analyzing parliamentary data", and "Experimental setup") and effectively identified significant events (see Figs. 2 and 4).

Sixthly, our study presents a generic framework with strengths and limitations in parliamentary data analysis. On the one hand, the framework adeptly captures and analyzes basic attributes such as protocol length and committee participation, offering valuable insights. However, introducing less generic features across different parliaments, such as nuanced details about specific parliaments' members or their voting patterns, poses a challenge for future exploration. In future research, we plan to extend the variety of features the framework supports.

Additionally, it is worth noting that as our dataset grows in size and complexity, relying solely on manual verification is becoming increasingly impractical. As a result, in future research, we intend to explore ways to incorporate automated event validation techniques into our methodology. One promising way involves harnessing the power of external sources, such as news channels or online social networks, that can effectively verify and confirm detected events.

We also showed that it is possible to utilize CPD algorithms, such as the PELT and the DYNP algorithms (see Sect. "Time series analysis"), to detect actual events that affected the parliamentary activities of different committees, such as summer breaks or the COVID-19 pandemic (see Fig. 4). We demonstrated that with some parameter algorithm tuning, such as time resolution tuning, it is possible to use CPD algorithms to identify events with relatively high precision. However, as shown in Table 6, different CPD algorithms can work differently: or example, the DYNP algorithm produced better results than the PELT algorithm yet there are also several cases where the PELT algorithm demonstrated higher precision than DYNP. Therefore, there is a need to develop a method to select the best performing CPD algorithm for each committee. Moreover, instead of utilizing only a single feature, it is possible to combine all features and different types of anomaly detection algorithms to detect abnormal events and obtain hopefully higher precision values. It is our hope to investigate this in future research.

Lastly, the methods and framework presented in this study have several limitations worth noting. As we demonstrated; the framework works in different parliaments. It still needs some manual adjustments to collect data from other parliaments. Furthermore, parliaments work in a number of different ways at different times of the year. Therefore, utilizing the various algorithms, such as IQR and CPD algorithms, may require manual adjustment. This includes time resolution setting and adjusting features calculations. Nevertheless, we developed our framework to be generic enough to support its extension to other parliaments.

Conclusions

This study presents the CAPD framework, an open and generic framework that enables collecting and analyzing large-scale parliamentary data from multiple sources. The framework can collect texts of parliamentary protocols, such as committee meeting protocols, and parse them into structured data (see Sect. "Methods and experiments" and Fig. 1), and analyze the data using state-of-the-art data science algorithms and tools. To test and evaluate the CAPD framework, we collected 64,913 protocols, from 264 committees, from three countries (see Sect. "Results"). We parsed these protocols and calculated different committees' features, and generated a time series illustrating the various committees' activities over long periods. Then, we utilized IQR to uncover abnormal events and CPD algorithms to detect events that influenced committees' activities (see Sect. "Methods and experiments"), such as the COVID-19 pandemic. Lastly, we evaluated the detected CPDs, which indicates that CPD algorithms performed adequately with relatively high precision rates.

Our research aims to expand the framework's reach to include additional countries like the UK. Combining advanced methods such as NLP with the ability to extract detailed features, enabling predictive analysis and comparisons between countries, computing complex features such as topics related to votes in parliament, thus facilitating predictive studies.

This study has several future research directions. First, we can expand the framework's collection capabilities and collect more data from more countries and other parliamentary bodies. Second, we can calculate additional features from

the protocols, such as the number of speakers and speakers' average word number. Moreover, we can combine the use of advanced methods such as Natural Language Processing (NLP) algorithms and tools, such as BERT Topics [38] and various sentiment analysis algorithms [39] to analyze the protocols and extract additional features regarding the topics and content of each protocol. Supporting this type of additional features will enable predictive analysis and comparisons among parliamentary bodies. Third, we can develop an interactive website that, based on CAPD framework analysis, presents different real-time features on other committees. The website can help the public monitor government activities better.

Lastly, the framework can be extended to collect data from different countries on specific topics. This data collection will enable comparative studies regarding governments' performance in handling global events, such as climate change and COVID-19.

Overall, the CAPD framework can harness worldwide open parliamentary data to advance government research and monitoring, promoting research in the field of public policy and helping make governments more transparent.

Acknowledgements
We thank Valfredo Macedo Veiga Junior (Valf) for designing the infographic illustration. We thank Polly Hember for proofreading this article. In addition, we would like to thank the anonymous reviewers for their careful reading of our manuscript and their insightful comments.

Author contributions
AM, SB, and MF conceived the study; AM and SB curated the dataset; AM developed the code framework; SB and AM evaluated the performances of the algorithms; MF supervised the study; and authors wrote the paper.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Data Availability
The CAPD framework's code and all the collected data will be available upon publication

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 29 May 2023 Accepted: 26 September 2023
Published online: 19 October 2023

References

1. Hulstijn J, Darusalam D, Janssen M, Baldoni M, Baroglio C, Micalizio R. Open data for accountability in the fight against corruption. In CARe-MAS@ PRIMA, 2017. pp. 52–66.
2. Kitsios F, Kamariotou M. Open data and high-tech startups towards nascent entrepreneurship strategies. In: Khosrow-Pour M, editor. Encyclopedia of information science and technology. 4th ed. Pennsylvania: IGI Global; 2018. p. 3032–41.
3. Fetisova OV, Kurchenkov VV, Golodova OA, Azmina JM. The role of information (smart) technologies in improving the efficiency of public administration. In: Institute of scientific communications conference. Springer. 2020. pp. 965–75.
4. Guggisberg S. Transparency in the activities of the food and agriculture organization for sustainable fisheries. Mar Policy. 2021;136:104498.
5. Ubaldi B. Open government data: towards empirical analysis of open government data initiatives. 2013.
6. Moaiad Ahmad Khder. Web scraping or web crawling: State of art, techniques, approaches and application. Int J Advn Soft Comput Appl. 2021;13(3):145–68.
7. Carrara W, Chan W-S, Fischer S, Steenbergen E. Creating value through open data: study on the impact of re-use of public data resources. European Commission. 2015.

8. Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government. Inf Syst Manag. 2012;29(4):258–68.
9. Martin S, Foulonneau M, Turki S, Ihadjadene M. Open data: barriers, risks and opportunities. In: Proceedings of the 13th European Conference on eGovernment: ECEG. 2013. pp. 301–9.
10. Zuiderwijk A, Janssen M, Choenni S, Meijer R, Alibaks RS. Socio-technical impediments of open data. Electron J e-Gov. 2012;10(2):156–72.
11. Foulonneau M, et al. Open data in service design. Electron J e-Gov. 2014;12(2):97–105.
12. Janssen K. The influence of the psi directive on open government data: an overview of recent developments. Gov Inf Q. 2011;28(4):446–56.
13. Ubaldi B. Rebooting public service delivery-how can open government data help drive innovation. 2016.
14. González-Zapata F, Rivera A, Chauvet L, Emilsson C, Zahuranec AJ, Young A, Verhulst S. Open data in action: initiatives during the initial stage of the covid-19 pandemic. 2021.
15. Huyer E, van Knippenberg L. The economic impact of open data: opportunities for value creation in Europe. European Commission, 2020.
16. Sohrabi C, Alsafi Z, O'neill N, Khan M, Kerwan A, Al-Jabir A, Iosifidis C, Agha R. World health organization declares global emergency: a review of the novel coronavirus (covid-19). Int J Surg. 2020;76:71–6.
17. Ibáñez L-D, Hoffman M, Walker J, Simplerl E. Sustainability of (open) data portal infrastructures a distributed version control approach to creating portals for reuse, 2020.
18. Laver M, Benoit K, Garry J. Extracting policy positions from political texts using words as data. Am Polit Sci Rev. 2003;97:311–31.
19. Awadallah R, Ramanath M, Weikum G. Opinions network for politically controversial topics. In: Proceedings of the first edition workshop on politics, elections and data, 2012. pp. 15–22.
20. Iyyer M, Enns P, Boyd-Graber J, Resnik P. Political ideology detection using recursive neural networks. In: Proceedings of the 52nd Annual meeting of the association for computational linguistics (Volume 1: Long Papers), Baltimore, Maryland, June 2014. Association for Computational Linguistics. pp. 1113–22.
21. Vilares D, He Y. Detecting perspectives in political debates. In Proceedings of the 2017 conference on empirical methods in natural language processing, 2017. pp. 1573–82.
22. Gencheva P, Nakov P, Màrquez L, Barrón-Cedeño A, Koychev I. A context-aware approach for detecting worth-checking claims in political debates. Proc Int Conf Recent Adv Natural Lang Process RANLP. 2017;2017:267–76.
23. Abercrombie G, Batista-Navarro R. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. J Comput Soc Sci. 2020;3:1–26.
24. Stavropoulou S, Romas I, Tsekeridou S, Loutsaris MA, Lampoltshammer T, Thurnay L, Virkar S, Schefbeck G, Kyriakou N, Lachana Z, et al. Architecting an innovative big open legal data analytics, search and retrieval platform. In Proceedings of the 13th international conference on theory and practice of electronic governance, 2020. pp. 723–30.
25. Cantador I, Sánchez LQ. Semantic annotation and retrieval of parliamentary content: a case study on the Spanish congress of deputies. In CIRCLE, 2020.
26. Varlamis I, Dalas A. Operational design and development of parliamentary recommender systems: the Hellenic parliament case study. Smart Parliaments. 2022. p. 35.
27. Porter MA, Mucha PJ, Newman MEJ, Warmbrand CM. A network analysis of committees in the us house of representatives. Proc Natl Acad Sci. 2005;102(20):7057–62.
28. Dal Maso C, Pompa G, Puliga M, Riotta G, Chessa A. Voting behavior, coalitions and government strength through a complex network analysis. PLoS ONE. 2014;9(12):e116046.
29. Barari S, Simko T. Localview, a database of public meetings for the study of local politics and policy-making in the united states. Sci Data. 2023;10(1):135.
30. Dorcas Wambui G, Waititu GA, Wanjoya A. The power of the pruned exact linear time (pelt) test in multiple change-point detection. Am J Theor Appl Stat. 2015;4(6):581.
31. Truong C, Oudre L, Vayatis N. Selective review of offline change point detection methods. Signal Process. 2020;167:107299.
32. Zeileis A, Shah A, Patnaik I. Testing, monitoring, and dating structural changes in exchange rate regimes. Comput Stat Data Anal. 2010;54(6):1696–706.
33. Erdman C, Emerson JW. A fast Bayesian change point analysis for the segmentation of microarray data. Bioinformatics. 2008;24(19):2143–8.
34. Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc. 2012;107(500):1590–8.
35. Scott AJ, Knott M. A cluster analysis method for grouping means in the analysis of variance. Biometrics. 1974;30:507–12.
36. Schwertman NC, Owens MA, Adnan R. A simple more general boxplot method for identifying outliers. Comput Stat Data Anal. 2004;47(1):165–74.
37. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognit Lett. 2009;30(1):27–38.
38. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.05794.
39. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-Based Syst. 2021;226:107134.

Publisher's Note