COPE
Member since 2019
JM14475

Discovering Representations of Democracy in Big Data: Purposive Semantic Sample Selection for Qualitative and Mixed-Methods Research

Hubert Plisiecki
SWPS University, Polish Academy of Sciences, Poland

Agnieszka Kwiatkowska
SWPS University, Poland; European University Institute, Italy

https://doi.org/10.18778/1733-8069.20.4.02

Keywords:
sample selection,
purposive sampling,
qualitative research,
word embeddings,
democracy

Abstract: The increasing volume of large, multi-thematic text corpora in social sciences presents a challenge in selecting relevant documents for qualitative and mixed-methods research. Traditional sample selection methods require extensive manual coding or prior dataset knowledge, while unsupervised methods can yield inconsistent results with theory-driven coding. To address this, we propose purposive semantic sampling – a Natural Language Processing approach using document-level embeddings created by a weighted average of word vectors with term frequency-inverse document frequency (tf-idf). We demonstrate its effectiveness using the example of democracy, a complex topic difficult to retrieve from parliamentary corpora. This method applies to any multi-thematic research area within big data, offering a reliable, efficient sample selection method for social research texts. Our contribution includes validating this NLP approach for social sciences and humanities as well as providing a robust tool for researchers, facilitating deeper qualitative analysis and exploration of big data corpora within the computational grounded theory framework.

Hubert Plisiecki
A PhD Candidate in Psychology at the Polish Academy of Sciences (PAN), works at the intersection of political science, psychology, sociology, linguistics, and machine learning. Granted MSc in Psychological Research Methods with Data Science. Research assistant at the Digital Social Sciences lab at PAN and in a project "Institutionalization of political parties in the parliaments of Central Europe – data mining of parliamentary debates" at SWPS University. His research interests include Machine Learning in Social Sciences, Meta-analytic Bias Detection.
e-mail: hplisiecki@gmail.com

Agnieszka Kwiatkowska
Assistant Professor at the SWPS University and Jean Monnet Fellow at the European University Institute, holds a PhD in sociology and an MA in political science. Her research focuses on political discourse: how issues are politicized, introduced into parliamentary competition, and become determinants of political behavior. Principal Investigator in the project "Institutionalization of political parties in the parliaments of Central Europe – data mining of parliamentary debates" (funding: National Science Centre), which investigates mixed methods of analyzing parliamentary speeches and voting.
e-mail: agn.kwiatkowska@swps.edu.pl

Introduction[1]

The digital age has ushered in an unprecedented volume of text data for social research, including data from social media content from platforms such as Twitter, Facebook and Instagram, forums and blogs, online news articles, books collections, and scientific literature. Additionally, because of the ongoing digitization of the public sector, many large datasets belonging to governments, public administration, and international organizations became available to the public (Schou, Hjelholt, 2018). In combination with rapid developments regarding computational power of computers and new Natural Language Processing (NLP) algorithms, which facilitate and accelerate the analysis of large datasets (Jemielniak, 2020; Foster et al., 2021), this wealth of data holds the promise of deeper insights into complex social phenomena; it also demands innovative approaches for efficient and accurate document selection.

In qualitative and mixed-methods research designs based on large, multi-thematic text corpora, the selection of the most relevant documents for analysis is a foundational step. While reducing the amount of data to be analyzed, it also greatly impacts the validity and reliability of research results (Krippendorff, 2018; Foster et al., 2021). Particularly in the era of the proliferation of big data, the need for effective identification of documents that are maximally pertinent to a research topic while avoiding intertwined irrelevant documents is widely called for (Wodak, Krzyżanowski, 2008; Krippendorff, 2018).

Methods of sample selection from big data currently used in social sciences are suboptimal. Traditional methods for document retrieval, such as keyword-based searches, have limited ability to accurately represent the content and the context of a document. Additionally, they require manual

[1] The research was supported by the National Science Center (research project 2019/33/B/HS5/02648).

browsing and document selection (Deterding, Waters, 2021; Saldaña, 2021), the usage of existing dictionaries or annotated corpora, e.g., The Manifesto Project (Lehman et al., 2024), or they are based on the differences in frequencies of keywords (Scott, Tribble, 2006). The main disadvantages of these methods are, respectively: high time consumption and labor intensity; the subjectivity of coders' interpretations; the need to limit the topics and scope of research to existing ready-made codes; and the possibility of selecting documents that are poorly-related to the research problem at the risk of omitting the important ones.

In recent years, the growing popularity of unsupervised methods, mostly based on latent topic modeling (Roberts et al., 2014), has eliminated restrictions resulting from labor intensity, coders' subjectivity, and limited availability of topic-specific dictionaries. Nevertheless, these methods are primarily aimed at the thematic classification of entire corpora, hence their use for retrieving documents most representative of the specific topic brings results that are incompatible with manual coding (Nicholls, Culpepper, 2021; Baden et al., 2022) as well as dependent on the prevalence of other topics in the corpus (Blei, Ng, Jordan, 2003).

In our article, we address these challenges by validating the NLP document retrieval method for selecting samples from big data in the social sciences and humanities. We demonstrate the effectiveness of purposive selective sampling by exemplifying its application through the sample selection parliamentary debates for the study of dimensions of democracy, which is a blurred concept, both regarding its meaning, dimensionality, and diachronic change, and highly overlapping with many other topics (Munck, Verkuilen, 2002; Coppedge et al., 2011). However, its application extends to any multi-thematic research area, offering a reliable and efficient method for data-driven sample selection of text documents in social research.

The main audience of this article comprises applied computational and qualitative social sciences and digital humanities researchers, for whom the use of the proposed method will facilitate the selection of the most relevant documents for their research from large collections of texts. Additionally, the article will be useful for NLP methodologists as an evaluation of the application of the existing retrieval methods. For this reason, we have included in the article a concise review of the most widely used methods of automated document selection, as well as the more detailed technical information necessary to further develop and refine this method. Our contribution lies, therefore, both in validating this NLP approach for social sciences and the humanities, and in providing a ready implementation of the proposed method (accompanying the article in the code repository), facilitating deeper and more extensive qualitative analysis and exploration of new big data corpora, particularly within the computational grounded theory framework (Nelson, 2020; Carlsen, Ralund, 2022).

Researching parliamentary debates

Parliamentary debates are a rich source of data for the social sciences and humanities, offering unparalleled insights into political discourse, policy evolution, and democratic practices. We have chosen to focus on parliamentary corpus analysis, from the numerous areas of the applicability of the proposed method, to address the growing importance and popularity of this approach among social and political scientists as well as machine learning and artificial intelligence researchers In recent decades, the availability of parliamentary materials in machine-readable formats has significantly increased (Rauh, Schwalbach, 2020; Erjavec et al., 2023). In addition, many parliamentary websites now offer search options based on keywords and metadata (e.g., author of the statement, political affiliation, time, subject) linked to speeches.

The availability and easy accessibility of parliamentary data is crucial not only to political analysts but also to non-governmental organizations and all citizens, helping them to hold politicians accountable by expressing democratic control over elected bodies (Lourenço, Piotrowski, Ingrams, 2017). Accountability, in turn, leads to better quality legislation (Voermans, ten Napel, Passchier, 2015) and lower levels of corruption (Lyrio, Lunkes, Castelló Taliani, 2018). Besides monitoring MPs' and governments' activities, transparency allows for wider inclusion of citizens in the process of democratic decision-making (Meijer, 2003) as well as increases support for representative democracy (Coffé, Michels, 2014).

However, the specific nature of parliamentary debates makes them more difficult for automated selection and analysis than most structured or semi-structured corpora. First, there are different types of debates in parliament, including reports on the activities of a ministry or government agency; interpellations and inquiries addressed primarily to government representatives; answers to questions; discussions on proposed legal regulations; debates on appointments and dismissals; adopting resolutions commemorating historical events; and, finally, debating the prime minister's exposé when forming a new government or a vote of confidence/no confidence in the government or individual ministers (Yamamoto, 2007; Ilie, 2015). Each of these types of debates has a different structure, length, number of speakers, interjections from different styles (e.g., extracts from legal texts or resolutions), or the presence of speakers from outside parliament (members of the government, the president, representatives of government agencies).

Second, in their speeches, MPs often express the interests of various entities: their own, their closest sociopolitical environment, their constituency, and their party. The speeches serve several purposes, including position-claiming, persuading, negotiating, agenda-setting, and opinion-building (Ilie, 2015). As a result, parliamentary speeches are characterized by a variety of styles and rhetorical figures, which further complicates their analysis. Third, MPs frequently reference previous or anticipated speeches in their statements, making joint analysis of the speeches necessary. Finally, interruptions from the floor, both anonymous and non-anonymous (Kwiatkowska, 2017) – and by the chairman of the session in the event of exceeding the time limit, violation of the rules of parliament, or as a tool to suppress political opponents (Shaw, 2000) – are not uncommon. Due to these idiosyncrasies, a single speech frequently contains a mixture of various topics, references to activities of other individuals and institutions, indirect addresses to voters, and formal parliamentary phrases. Thus, while the accessibility of parliamentary data has greatly improved, the sheer volume and complexity of this data poses significant challenges for extracting relevant documents for in-depth qualitative analysis, leaving much of this valuable resource under-researched and underlining the need for the advanced document retrieval method.

Democracy as a complex and multifaceted concept

In addition to the analytical difficulties caused by the specificity of parliamentary debates, the quality of the retrieval of documents for a specific topic largely depends on the exclusivity and coherence of this topic (Bischof, Airoldi, 2012). In some cases, the problem of document selection is easily solved, particularly when the researched topic is very specific and isolated (unrelated to other topics or related only in rare, known cases), and when it appears in a limited number of documents. However, in the case of complex and diffused research topics, the correct selection of relevant documents (or parts of them) is a difficult problem that significantly impacts the quality of research results (Wodak, Krzyżanowski, 2008).

The concept of *democracy* stands as a perfect example of the conceptual complexity confronting social researchers, being a multifaceted topic, compounded by varying interpretations across contexts. Despite numerous studies in social and political sciences, no scientific consensus has been reached on the content of democracy and its attributes and types, and it remains a highly contested issue to this day (see: Cunningham, 2002; Munck, Verkuilen, 2002; Coppedge et al., 2011). While there is agreement among researchers regarding the core elements of democracy – including free and fair elections, civil and political liberties, and government accountability – the scientific debate on democracy mainly concerns dimensions such as political inclusion (Dryzek, 1996; Young, 2002), economic egalitarianism (Boix, 2003; Knutsen, Wegmann, 2016), and social equality, including ethnic equality (Houle, 2015) and gender equality (Beer, 2009). Furthermore, the public understanding of democracy has been shown to vary across time and space (Dalton, Shin, Jou, 2007; Ferrín, Kriesi, 2016), among age groups (Sack, 2017; Nieuwelink, ten Dam, Dekker, 2018), due to the socio-economic status (Ceka, Magalhães, 2020), and in response to political regime transition (Dalton, Shin, Jou, 2007; Sack, 2017).

Additionally, the concept of democracy differs significantly from the regular policy topics in parliamentary debates. Except during early democratic transition periods, when the foundations of the political system are being established (Elster, Offe, Preuss, 1998), or in the event of a threat to the democratic system (Levitsky, Ziblatt, 2018), the topic of democracy is rarely the subject of parliamentary discussion *per se*. Rather, it frequently appears as a meta-issue, in the form of distant references loosely-related to the topics currently being discussed or as the benchmark by which public policy solutions may be judged on their compliance with democratic principles. Alternatively, references to democracy are used as a political weapon to discredit political opponents (Kwiatkowska, Muliavka, Plisiecki, 2023).

The multiple meanings and scattered references to democracy in parliamentary debates make it challenging to locate relevant documents. Retrieving documents by searching for a single keyword that represents the topic is inadequate due to its vague and noisy nature. Manually analyzing each speech in a large corpus is also impractical. While for some topics, there are manually-annotated corpora (e.g., Lorenzini et al., 2022) and domain-specific dictionaries (e.g., Albaugh et al., 2013), none exist for democracy. Albaugh creating such a dictionary is possible (e.g., on the basis of Varieties of Democracy classification; see: Kwiatkowska, Muliavka, Plisiecki, 2023), it is time-consuming, prone to subjectivity, and limited in scope due to the topic's complexity and variability. Keyword methods (see: Scott, Tribble, 2006) – which compare the frequencies of words in the corpus to similar text collections (reference corpora) – are also insufficient for dealing with unknown categories and the lack of human assistance in topic classification.

Computational purposive sample selection from large corpora

In the last two decades, automatic methods of pattern recognition – which is the prerequisite for further non-random, purposive sample selection – have been increasingly used in the social sciences and humanities (Foster et al., 2021). Of particular importance is the emergence of the computational grounded theory field (Nelson, 2020; Carlsen, Ralund, 2022), employing machine-learning methods to enhance the analysis of qualitative data. Unsupervised methods of natural language processing can be interpreted as automated implementation of grounded theory (Glaser, Strauss, 1999). They do not impose top-down interpretations on the researcher through a pre-developed theoretical model, allowing the themes to emerge inductively from data. Thus, not only can they manage extensive datasets that would be overwhelming for manual analysis, but they also remove the problem of subjectivities, which can influence both data analysis and collection (however, see Charmaz, 2006 on managing subjectivity).

Latent Topic Modeling methods

Particularly widespread in qualitative and mixed-methods studies based on large data collections is latent topic modeling (LTM; see Kherwa, Bansal, 2019 for a comprehensive review) – a class of unsupervised methods classifying documents into semantically-interpretable representations of distributions based on frequency of co-occurring (Roberts et al., 2014), enabling researchers to uncover thematic structures within datasets. While LTM eliminated restrictions in social research resulting from the labor intensity and the availability of data, it became apparent that the results of fully data-driven unsupervised models are inconsistent with the results of manual coding, which emphasizes theory-driven topics (Baden et al., 2022; Nicholls, Culpepper, 2021). To address this issue, enhanced, semi-supervised versions of topic models were developed using an already existing training set of annotated documents (Blum, Mitchell, 1998) and publicly available knowledge repositories, such as Wikipedia (Schönhofen, 2009). More advanced models combine topic modeling with prior sets of words to train document classifiers (Watanabe, Zhou, 2022). Unfortunately, due to the lack of coded collections, they are not usable in most research topics, including on democracy.

Second, as LTM methods were created with exploratory analysis in mind, their usage to find specific, predetermined topics is problematic, as it has to be done either through an iterative process of rerunning the topic model with different parameters or by feeding it a predetermined number of seed words (i.e., initial reference points or categories that help guide the learning process of the model when initial labels or categories are not predefined and allow to replicate the analysis). While both resolutions are feasible at first sight, each of them has unforgivable flaws. Namely, a model trained to recover multiple topics at the same time is redundant when only one topic is of research interest. Furthermore, the probability distribution of that one topic is affected by the probability distributions of the rest of the topics recovered by the model (Blei, Ng, Jordan, 2003). On the other hand, seed-word-based models need the collection of seed words based both on the theoretical assumptions and the distribution of word tokens already present in the text (Wang, Thint, Al-Rubaie, 2012). Given the unreliability of LTM methods for determining document relevance in large corpora studies, there emerges a clear need for a more refined qualitative sample selection method in large corpora studies. Building upon the advances made within the NLP field, we validate for social sciences the application of first-stage answer retrieval algorithms for quantifying document relevance for a given topic of interest.

The word-embedding-based methods

Searching for documents related to a query of interest is an old and widely researched topic in computer science. Modern search systems usually approach this problem by partitioning the task into two subtasks: first- and second-stage retrieval. The goal of the former is to retrieve a large number of candidate documents from the corpus to be later ranked using the more precise second-stage retrieval algorithm (see: Guo et al., 2022). The current study is concerned with the former since its goal is to identify a wide range of documents which are relevant to the studied topic.

The first-stage algorithms come in different forms, starting from the oldest, relying on word frequencies, to contemporary neural net-based approaches (Guo et al., 2022), which are able to create meaningful representations of documents by capturing term order and contextual relations between distant terms. However, transformer models (e.g., BERT; Devlin et al., 2019) and deep learning methods require a lot of resources to implement, which might not be available to social scientists. While open-source pre-trained models exist (Wolf et al., 2020), their implementation is still expensive in terms of processing time and memory, and requires advanced programing skills. Furthermore, due to their pre-trained nature, it is not certain that the corpora on which they have been trained are suitable for the idiosyncratic texts studied in social sciences.

Currently, yet another way of retrieving text is available, namely making use of Open AI embeddings, which has proven performance (Xian et al., 2024) or other API's that provide high quality vector embeddings. This option, while easy to implement, can be resource-intensive, especially when working with really large datasets. The use of such tools is, however, restricted only to those that are willing to learn how to use the application programing interface of the platform they wish to take their embeddings from.

For these reasons, we propose using cheaper, less complex, yet robust techniques of term dependency models, known as word embeddings. Word-embedding models are created by reducing the dimensionality of a term co-occurrence matrix to create a dense numerical representation of textual data (Mikolov et al., 2013a). These models assign vectors (strings of numbers) to each word in the corpus to represent its meaning based on its co-occurrence with all other words in the text. The similarity (distance) between words can then be computed using the vectors, which exist in an embedding space where similar word embeddings are close to each other and different ones are far apart.

Word-embedding models have two main weaknesses: 1) vocabulary mismatch (i.e., a mismatch between the available numerical representation and the corpus of interest); and 2) the inability to capture term-ordering (word order) (Guo et al., 2022). These issues are critical in document retrieval in computer science, where retrieval must be performed dynamically on novel texts. As more data is introduced, a set of word embeddings trained on a specific primary corpus of documents can lose its relevance, rendering word vectors calculated for certain terms increasingly useless as the meaning of the term evolves over time. Additionally, information retrieval in computer science aims to deliver the exact specific information that the user requests while ignoring the wider semantic context.

Fortunately, both of these obstacles become less significant in social sciences and the humanities when performing document retrieval for the purpose of selecting a diverse set of documents for further qualitative research. In the majority of social studies, unique numerical representations can be created for each corpus, ensuring that the vocabulary of the representation model matches that of the corpus at hand and avoiding obsolescence as the corpus grows (Guo et al., 2022). As the volume of data increases, new representation models can be retrained on the enlarged corpus. Additionally, when conducting qualitative research, researchers are interested in a broad range of documents related to the topic rather than a single document with the most precise answer. Therefore, the importance of term order or adjacency in the query is reduced, and a cruder method of locating relevant documents is sufficient.

Data and methods

Background

Based on prior research in information retrieval, we utilized word embeddings to locate a predefined topic in a large corpus of speeches. We tested four different methods, with gradually increasing complexity, alongside a baseline term-based method, and validated their effectiveness for qualitative research. Unlike in computer science, where accuracy is typically used to evaluate document retrieval methods (Gillick, Presta, Tomar, 2018), qualitative research prioritizes identifying a diversity of forms and contexts in which a topic appears in the corpus rather than the precise accuracy of the classification of each document (Patton, 2015). While the two goals overlap, the former one is more related to recall, while the latter is more related to precision. Taking this into consideration, we shifted our focus from accuracy metrics to qualitative ones, evaluating the coherence and exclusivity of topic models applied to documents selected using those methods.

Datasets

The proposed retrieval methods were tested on the Polish parliamentary corpus. The full corpus of debates from the official proceedings of the lower chamber of the Polish parliament (*Sejm*), spanning 1991 to 2020, was extracted from its website (http://www.sejm.gov.pl) and comprised over 291,000 plenary speeches. After extraction, the corpus was cleaned of structural errors and the speeches were morphologically analyzed using the Morfologik software, version 2.1.6 (Miłkowski, 2022), to track multiple forms of the words simultaneously. Stopwords were removed at each step of the analysis.

Word embeddings

The embeddings were created using the python package *gensim* implementation of the word2vec algorithm with negative sampling (Mikolov et al., 2013b). The size of the embeddings was set to 300, and the default hyperparameters were used for training the main word-embedding model. Subsequently, for training the embedded topic model, the hyperparameters were modified with alpha set to 0.03, sample to 6e-5, min_alpha to 0.0007, and negative value set to 20. Two stages of preparing the word embeddings are the result of, first, using the entire corpus for document retrieval and creating five subcorpora in all methods except the count method, and, second, evaluating the selected five subcorpora using the embedded topic model (for all methods).

Topic retrieval methods

Four different methods of document retrieval based on word embeddings are proposed, alongside a scaled count term method which serves as the baseline. The methods gradually increase in complexity to demonstrate the advantages of using more advanced approaches.

The count method

The first method is the most rudimentary and assumes that if a document contains one or more instances of the word “democracy”, it is relevant to that topic. To test this approach, we compute the number of times the stem of the word “democracy” (*demok* in Polish, En. *democ*) appears in each document. The resulting sum is divided by the number of words in the document to control for its length. This is the baseline for the other methods as it does not make any use of word embeddings.

The ten-words method

The ten-words method assumes that only certain words in a document are relevant to the studied topic. We begin by computing the cosine similarity, which approximates the similarity of the words related to these embeddings, of all words in each of the documents. Then, we select ten words with the highest cosine similarity to the word “democracy”. Finally, we sum their cosine similarity values and divide the result by the number of words in the document to counteract the fact that longer texts have a higher chance of including words closely connected to the word “democracy”, because they have more words overall.

The separate-words method

The third method relies on the word embeddings created for the entire corpus. For each word in the document, we compute the cosine similarity of its embedding with the embedding of the word “democracy”. Therefore, if a document has many words that are similar to the word “democracy”, it should be relevant to the topic of democracy as well. We add up the similarity metrics of all the words in a document and divide the result by their number to normalize the metric and avoid longer documents producing higher similarity results simply because they contain more words.

The normal-mean (norm-mean) method

This approach uses numerical representations of entire documents instead of separate words. We average the word embeddings of individual words to produce document-wide embeddings.

The tf-idf mean method

The final method takes into account the importance of each word in a document with the use of tf-idf scores, which enhance the quality of produced document embeddings (Schmidt, 2019; Meijer, Truong, Karimi, 2021). These scores quantify how common a word is in a particular document alongside how unique it is in the entire corpus. We use them to compute a weighted average of the words in each document, and this ensures that documents with unique content retain this unique information in the process of being transformed into document embeddings. The tf-idf weights apply higher weights to unique information within documents, making it less likely to get lost in the process of compressing the word embeddings into document embeddings.

The selection of sub-corpora for methods’ evaluation

All methods were evaluated both qualitatively and quantitatively. However, before conducting analysis, it is crucial to highlight one methodological caveat. For a fair comparison of techniques, documents should be selected from the corpora based on their relevance scores from the applied methods, creating a separate sub-corpus for each method. To ensure comparability between the sub-corpora, the number of documents in each must be kept constant. Therefore, for evaluative purposes, we selected the highest-ranked documents from each method. Their number was determined by the number of documents with a non-zero score returned by the count method (all documents mentioning democracy at least once) – in our case, 15185. Accordingly, this number of relevant documents was selected for each method, resulting in five different sub-corpora. However, for further qualitative research, a much smaller subset (e.g., 100 documents with the highest scores) may be more practical. The outcomes of our analysis are detailed in the next section.

Results

After selecting the sub-corpora, we inspected their overlap by employing both the Venn diagram and the quantitative approach of computing the correlations between the relevance scores of the documents within the shared set. The cross-corpus relationships depicted in Figure 1 demonstrate that the sub-corpora selected through the ten-words method (comprising 10,112 or 66% unique documents, not selected by any other method) and the count method (comprising 8,669 or 57% of documents specific to that corpus) notably differ from the rest. Conversely, the three sub-corpora derived through word-embedding techniques applied to entire documents (tf-idf mean, norm-mean, and scaled separate words) display a considerable overlap, indicating a high level of similarity in the documents they selected.

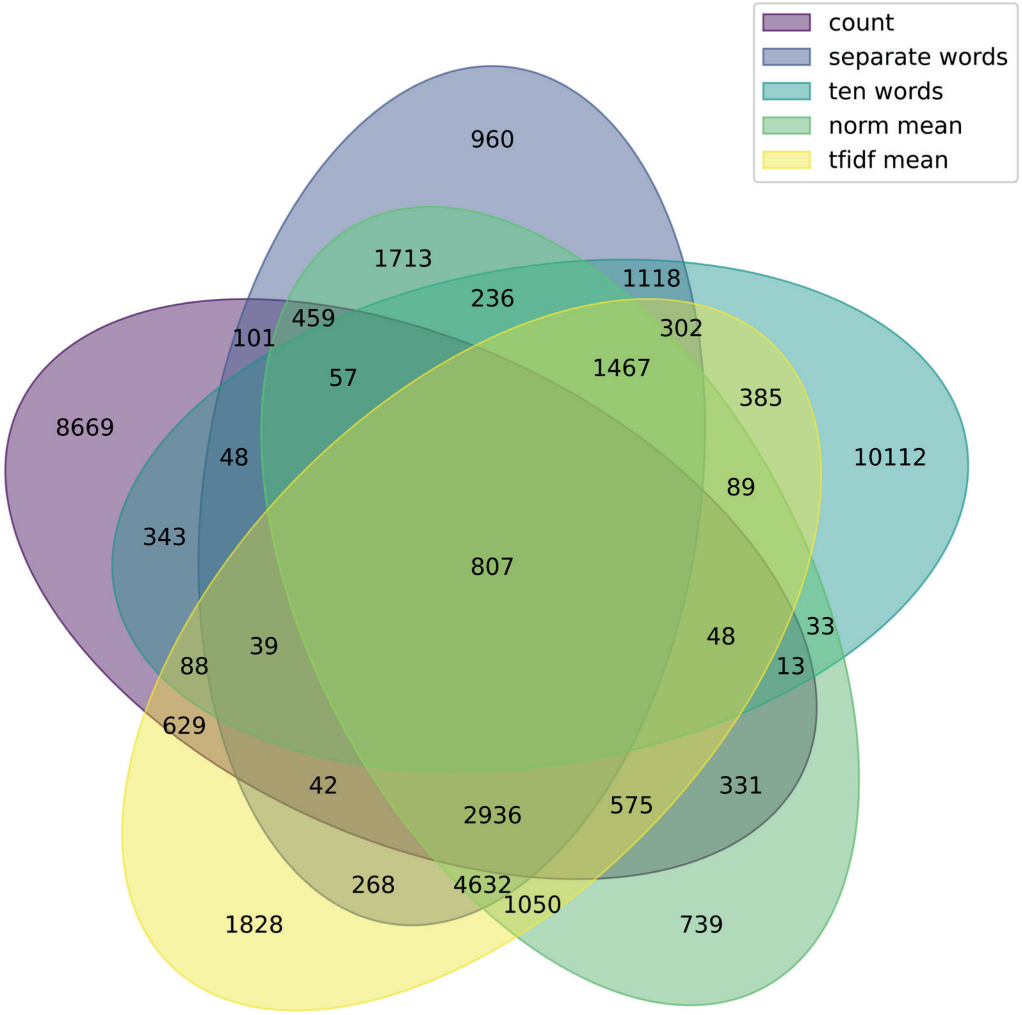
Further examination of the similarity in the documents' relevance scores across various methods for sample selection revealed that the three methods that utilize word embeddings for whole documents – tf-idf mean, norm-mean, and scaled separate words – show a medium-to-high strength of correlation, with correlation coefficients ranging between 0.52 and 0.74 (Figure 2). In contrast, the count method displayed a moderate correlation with the other methods, displaying coefficients between 0.20 and 0.56. The ten-words method exhibited the lowest correlation with other methods, particularly those employing word embedding on entire documents.

To examine which method most effectively captures the essence of democracy – namely by generating a broad range of topics that are directly relevant to the research problem – we applied topic models to each of the separate sub-corpora. For this purpose, we used the embedded topic model (ETM; Dieng, Ruiz, Blei, 2020), which is different from more traditional topic models such as LDA by requiring pre-trained embeddings for topic identification. To accommodate the unique characteristics of each method's selected documents, we trained specific embeddings for each sub-corpus. Each ETM model was configured to identify 20 topics, operating with a learning rate of 0.0005 over 1,000 epochs, thus providing a comprehensive exploration of how democracy is represented across the variously retrieved document sets.

The first step in further analysis of topics involved refining the set of topics by eliminating those that were either unintelligible or concerned with the formalities of parliamentary proceedings.

Subsequently, we named the remaining topics according to their thematic content and analyzed the results using quantitative and qualitative methods. For the quantitative analysis, we applied a series of metrics to evaluate the results of the embedded topic modeling for each sub-corpus.

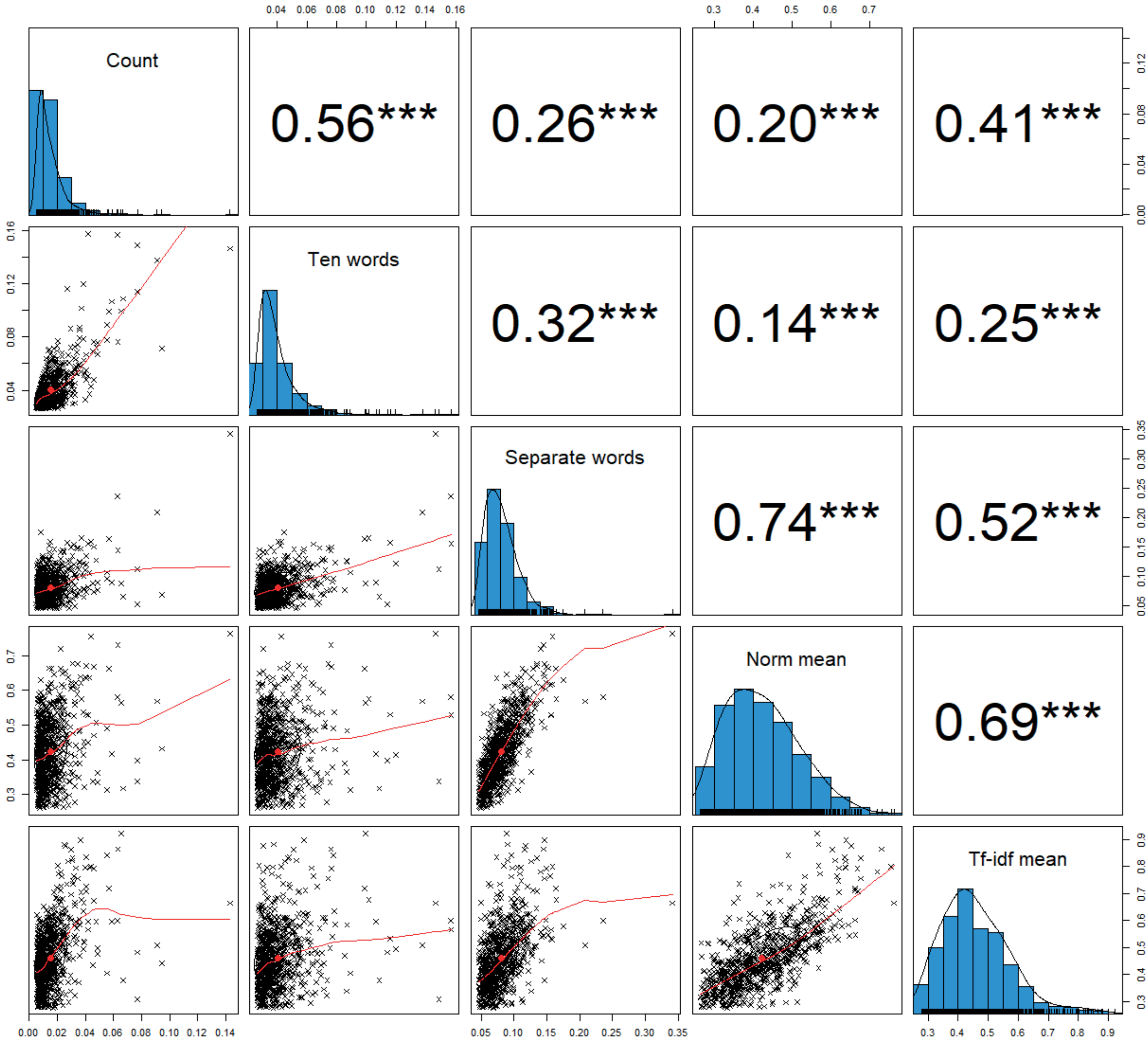
Figure 1. The Venn diagram of sub-corpora selected by five retrieval methods



Source: own calculations.

The first two metrics in Table 1 adhere to the standards set forth in the seminal work on the Embedded Topic Model (Dieng, Ruiz, Blei, 2020). *Coherence* measures the likelihood of words within a topic co-occurring in documents, serving as an indicator of topic consistency. *Diversity* is a metric of the distinctiveness of topics, calculated as the percentage of unique words within the top 25 words across all topics. Higher coherence and diversity indicate higher quality of the recovered topics. Additionally, we calculated the *Vocab length*, which represents the number of unique tokens extracted from the selected documents for model training. The average length of documents (*Mean document length*) was used to assess the representativeness of the documents selected by each technique compared to the whole corpus.

Figure 2. The correlation matrix of the document-level relevance scores in the shared subset



Note: Spearman's correlation coefficients between the documents' relevance scores assigned by each of the five methods on the shared subset of 807 documents.

Source: own calculations.

The evaluation of the metric reveals that the ten-words method falls behind the other techniques, generating substandard topics. Its topics registered the lowest scores in both coherence and diversity across all evaluated methods. The differences in coherence and diversity scores among the remaining methods were negligible. The ten-words method also had a significantly limited vocabulary, comprising only 6,663 tokens, whereas the vocabularies of other methods ranged from 17,814 tokens for the count method to 19,318 tokens for the norm-mean method, indicating a more extensive lexical variety. Furthermore, the average length of documents retrieved by the ten-word method was notably shorter, at just 67 words, falling substantially short of the corpus' average document length of 409 words. In contrast, the documents selected using the count method were on average twice the length of the corpus average, demonstrating its tendency to favor longer documents. The embedding-based methods, particularly the separate-words method, approximated the corpus average most closely.

Table 1. Comparative metrics of topic models generated using different selection methods

Method	Count	Ten-words	Separate words	Norm-mean	tf-idf mean
Coherence	0.20	0.13	0.20	0.20	0.20
Diversity	0.87	0.80	0.87	0.88	0.87
Vocab length	17,814	6,663	18,520	19,318	18,430
Mean document length	913	67	397	502	480

Note: Quantitative metrics calculated for each of the five methods (in columns). Coherence and Diversity are metrics proposed in the original ETM article (Dieng, Ruiz, Blei, 2020). Vocab length refers to the number of tokens that were retrieved from the selected documents and used to train topic models. Mean document length is the average number of words in the documents selected by each method.

Source: own calculations.

For the qualitative analysis of the topics, we examined the composition and content of each topic returned by the five methods. This involved examining the words and phrases most strongly associated with each topic to understand its essence. Then, we determined the number of coherent and relevant topics produced by each method and identified topics that were common across different methods using topic summary tables. Beyond the quantitative metrics of model quality presented in Table 1, we also analyzed the thematic coherence and specificity of each topic qualitatively. Finally, we discussed the overall quality of each model, taking into account both quantitative and qualitative insights.

Overall, as shown in Table 2, there is a considerable degree of thematic overlap among the topics identified by the ETM across the five distinct sub-corpora. The variation in topics' composition across different methods appears to mirror the correlation patterns observed in relevance scores, as illustrated in Figure 2. Notably, several topics emerged consistently across all sub-corpora, encompassing areas such as *EU/Foreign policy*, *Elections*, *Budget*, *Law and courts*, and *Family and social services*. These recurring themes underscore the methods' collective ability to capture key aspects of the dataset, despite their varied approaches to document retrieval and topic generation.

Table 2. The topics generated by the ETM on different corpora

	Sample selection method				
Topic	**Count**	**Ten-words**	**Separate words**	**Norm-mean**	**tf-idf mean**
EU/Foreign policy	European, union, Europe, foreign, country, politics, international, treaty, security	Polish, European, country, union, republic, year, Europe, agreement, foreign	European, union, Polish, politics, Europe, foreign, country, treaty, cooperation	European, union, Europe, politics, foreign, country, treaty, cooperation, international	European, union, Polish, politics, Europe, foreign, country, treaty, cooperation
War & independence	history, war, solidarity, martial, independence, memory, PRL, death, victim		Polish, Pole, nation, great, national, world, country, history, solidarity	war, soldier, nation, fight, Warsaw, military, global, victim, German	Polish, year, nation, Pole, time, world, great, history, fight
Public media	public, service, control, television, office, information, media, security, police		public, council, television, national, media, information, chairman, activity, professor	council, public, television, court, national, media, judge, justice, information	council, public, television, media, national, culture, radio, program, 20
Elections	electoral, elections, party, referendum, vote, majority, voting, choice, candidate	Sejm, proposal, voting, elections, electoral, point, debate, order, formal	project, act, elections, electoral, party, change, propose, amendment, ordination	elections, electoral, party, referendum, vote, change, parliament, local government, voting	electoral, elections, change, party, referendum, political, local government, small, voter
Local government	local government, municipality, thousand, million, city, voivodeship, territorial, local, agriculture	social, local government, program, resource, society, organization, possibility, municipality, assistance			
Budget	budget, tax, budgetary, zloty, finances, bank, fund, income, economy	year, budget, million, zloty, money, thousand, Pole, tax, road	government, prime minister, economy, money, economic, politics, coalition, program, budget	money, economy, ownership, billion, cost, zloty, farmer, assets, privatization	economic, economy, budget, reform, program, money, politics, level, million
Law & courts	law, tribunal, constitutional, court, proceeding, regulation, judge, justice, person	law, constitution, constitutional, article, principle, tribunal, court, judge, justice	law, constitution, constitutional, act, tribunal, court, legal, article, principle	law, constitution, constitutional, act, tribunal, article, legal, principle, regulation	constitutional, constitution, tribunal, court, act, article, judge, principle, regulation

Civil rights	law, political, citizen, authority, life, human, democratic, freedom, public				law, citizen, freedom, person, civic, protection, spokesman, legal, public
Family & Social services	person, work, child, family, school, vocational, health, worker, social	person, human, child, life, school, family, solidarity, teacher, young	human, life, law, child, family, freedom, woman, citizen, security	life, law, human, freedom, child, family, world, woman, value	life, human, child, family, woman, solidarity, church, value, young
Political system		state, political, citizen, PiS, democracy, own, party, human, self	state, political, democracy, action, country, system, create, society, democratic	state, issue, action, national, country, union, decision, nation, democracy	state, political, action, own, country, social, important, large, society
PiS government				power, PiS, service, to rule, state-owned, type, serve, take, own	PiS, newspaper, head, against, lustration, lie, protest, hour, eye
National memory			memory, anniversary, independence, holiday, fatherland, priest, death, august, victim	fatherland, tradition, anniversary, history, soil, holiday, saint, independence, generation	place, freedom, event, anniversary, Warsaw, city, holiday, saint, independence
PiS traditional values			PiS, value, culture, church, school, word, language, respect, tradition		
Security			state, service, operate, action, security, head, defense, prosecutor, police		authority, day, republic, national, person, force, defense, recall, service
Number of relevant topics	9	7	11	10	12

Note: The table displays the categories of topics recovered by each method, classified based on their keywords.

Source: own calculations.

Other topics that were frequently returned by the models were: *War and independence*, *Public media*, and the *Political system*. On the other hand, when looking for the least prevalent topics, only one topic, *PiS traditional values* (being a concoction of the topical Family and social services with references to Law and Justice and to national tradition) was exclusively recovered from a single sub-corpus using the norm-mean method. Several topics were present in only two corpora, including *Local government* (although some mentions of this topic were also found in the *Elections* topic returned by norm-mean and tf-idf methods), *Civil rights* (falling close to *Family and social services*), *PiS government*, and *Security*, and one topic was returned by three methods (*National memory*). In addition, for more specific and isolated topics, such as *Law and courts*, all methods retrieve highly-relevant legal terms, though the depth and specificity vary.

Overall, only seven intelligible and relevant policy topics were identified with the ten-words method, compared to 12 topics being identified with the tf-idf method and 10 to 11 with the other methods. Regarding topic specificity and depth, the count and ten-words methods often return less comprehensive term lists compared to other methods, suggesting a possible limitation in capturing a broad spectrum of relevant terms. Methods such as tf-idf mean and norm mean tend to offer more consistent and comprehensive lists of topics, indicating higher effectiveness in capturing diverse and relevant contents.

Discussion

Beginning with the most pronounced issues, the ten-word method notably underperforms in both quantitative metrics and qualitative analysis. It over-samples very short texts, leading to a limited vocabulary range, which, in turn, adversely affects both diversity and coherence metrics. A possible cause for this limitation is the extensive normalization applied to the method. Dividing the relevance score by the total number of words in the entire speech instead of just ten words overshadowed the actual value of the similarities between words themselves. Given its overall sub-par performance, we will not focus further on this method in this discussion.

Next, we will comment on the correlations between the different methods shown in Figure 2. The count method shows a moderate correlation with all other methods, serving effectively as a baseline while also suggesting potential for improvement, showing that the other proposed techniques can significantly differ from it. The similarity of the highly-intercorrelated embedding-based techniques stems from the fact that they attend to all words in each document, but differ in how they weigh these words. The more different the weighting scheme, the less similarity is observed between them, particularly noted in the lower correlation between the tf-idf mean method and the separate-words methods.

The average document length also merits attention. Assuming the null hypothesis of no difference in length between documents related to democracy and those in the general corpus, the most representative method should return documents whose mean length is close to the corpus average. However, the count method retrieved documents more than twice the average length. Interestingly, this tendency did not significantly impact the overall vocabulary length, which was still shorter than that retrieved by embedding-based methods. This suggests that while the count method gathers a higher number of words per document, it captures repeated rather than unique words.

Since the norm-mean, tf-idf mean, and separate-words methods are highly intercorrelated, their qualitative performance is similar regarding topic composition and content. In most cases, if a topic appeared in any one of these sub-corpora, it also appeared in the others. Unlike these three methods, the count and ten-word methods utilize only a limited number of specific words within each document. Consequently, they are more likely to select documents where the word "democracy" appears incidentally rather than as a central theme.

The three methods based on word embeddings, on the other hand, make use of all words in each document, enhancing their sensitivity to the entire document's relevance to the theme of democracy. This is exemplified in the topics selected by at least two methods from this trio but missed by the count and ten-word methods, including *PiS government*, *National Memory*, and *Security*. All of them are closely related to the issues of democracy in Poland, particularly to the process of democratic backsliding deepening since 2015, authored by the conservative and nativist PiS government in the period of 2015–2023 (Levitsky, Ziblatt, 2018).

When assessing these three embedding methods against each other based on quantitative criteria alone, it is difficult to choose the best one due to their similar output. However, adding a qualitative analysis of the results suggests that the tf-idf method yields the highest number of the most coherent and understandable topics compared to the separate-words method, omitting the topic related to *PiS traditional values* (returned only by the separate-words method) as well as returning highly-relevant topics of *Civil rights*, *Security*, and *PiS government*. This, combined with quantitative metrics from the information retrieval field (Gillick, Presta, Tomar, 2018), favors the tf-idf method.

Conclusions

The aim of the study was to determine the optimal – i.e., effective, yet and at the same time reasonably easy to use – method of selecting a sample from any large corpus of text documents for the qualitative or mixed-methods research in the social sciences and humanities. We proposed a purposive semantic sampling method based on word embeddings, which offers a substantial improvement over traditional keyword-occurrence counting. This method provides a continuous relevance score for each document, thereby allowing researchers to tailor the sample size to their specific needs. Given the increasing volume of textual data available for research, this approach exemplifies an effective big data strategy that both enhances the qualitative analysis and significantly reduces the amount of manual work required, thus saving researchers' resources.

We evaluated the effectiveness of this method using as the example the sample selection of documents most related to the concept of democracy from a corpus of Polish parliamentary speeches. Democracy, with its inherent ambiguities, served as an ideal subject to highlight both challenges encountered during the sample selection process and the strengths of our approach. We showcased the primary strength of the method presented in this article, as it does not require an extensive list of keywords, the creation of which is difficult and time-consuming, as well as prone to bias. As the method performed very well with such a complex and "noisy" meta-topic, we can conclude that it should be highly effective for simpler (i.e., compact and limited) research topics. On the other hand, it is in the case of multifaceted and difficult to retrieve topics that the proposed method is the most useful. Albaugh the embedding methods are portrayed as consistently better than the word-frequency methods in the document retrieval field (Guo et al., 2022), the count method may still prove to be of sufficient quality in certain cases, especially when the studied topic is narrowly-defined and isolated from other topics.

Our study demonstrated that using document embeddings enriched with tf-idf weights can improve the quality of sample selection and is, overall, a good choice for extracting documents that are representative and meaningfully related to the topic of interest. Moreover, it effectively manages synonyms and variations in terminology, thereby broadening the scope of document retrieval beyond exact query matches. In addition, the proposed method promoted the maximum variability, capturing the highest number of topics representing a wide array of narratives and themes within a general subject of democracy. Therefore, it aligns closely with the logic of qualitative, purposive sampling, which aims to capture information-rich cases closely related to the studied topic and representing a diversity within the population of interest rather than achieving empirical generalizations (Patton, 2015: 401).

Finally, beyond selecting a sample of the most relevant documents, the proposed method returns word embeddings for documents which can be used to further inform qualitative analysis by discovering underlying topics grounded in data and clustering documents into topics. While avoiding over-reliance on quantitative methods, which might lead to a lack of depth of data interpretation, word embeddings can capture semantic relationships in data that might not be evident through traditional coding or thematic analysis alone. Thus, referring to the illustrative example of the understanding of democracy analyzed in the article, obtained results complement using other empirical methods, including social surveys among general population (e.g., Ferrín, Kriesi, 2016; Ceka, Magalhães, 2020; Kwiatkowska, Grzybowska-Walecka, 2024) and political elites (Katz, 2001; Markowski, Kwiatkowska, 2018), expert-based international indices (Freedom House, 2022; Varieties of Democracy, 2022), as well as quantitative analyses of political texts (Kwiatkowska, Muliavka, Plisiecki, 2023; Schwörer, Koß, 2023).

Limitations and future work

The main limitation of the study arises from the specificity of the corpus used to perform the evaluation, i.e., parliamentary speeches. Its unique characteristics may restrict the generalizability of our results across different corpora and topics within the social sciences and humanities. However, given the increasing popularity of analyses of parliamentary speeches in social-sciences research (see: Back, Debus, Fernandes, 2021), this validation carries significant value. Further work is needed to assess the representativeness of our validation for the broader range of corpora and themes. Additionally, despite the specificity of formal and linguistic features of parliamentary debates, there is nothing inherent to the presented methods that could significantly affect the results of research conducted in various types of corpora and in different fields.

However, some technical issues are present. The query used in the implementation was simple, as it consisted of one word and, therefore, was represented by single-word embedding. For longer queries, we propose that researchers first convert the multi-word query into separate embeddings and then average them to create one embedding used for retrieval. This approach, however, may reduce precision, particularly for longer queries, whose embeddings might skew the overall meaning of the query. Addressing this challenge might involve assigning differential weights to the query's components or simplifying the query to essential keywords that capture the topic's essence without adhering to grammatical correctness. Future work, aside from testing the feasibility of various query formulations as discussed above, can extend the methodological framework proposed in this article to validate more advanced techniques of document retrieval applicable in the social sciences and humanities, such as those relying on advanced deep learning (Guo et al., 2022).

To facilitate broader application of the proposed document-level embeddings incorporating tf-idf weights as the purposive semantic sample selection method for qualitative and mixed-methods studies based on large corpora, we are releasing the Python package *retfidf* (Plisiecki, 2024) alongside this article. This package allows for the application of the tf-idf method in a user-friendly manner, making it accessible even to those with limited technical expertise in programing. In addition, we provide a dataset of parliamentary speeches on democracy (Plisiecki, Kwiatkowska, 2022) used in the article.

References

Albaugh Quinn, Sevenans Julie, Soroka Stuart, Loewen Peter J. (2013), *The Automated Coding of Policy Agendas: A Dictionary-Based Approach*, "Paper presented at the 6th Annual Comparative Agendas Conference", Antwerp, Belgium, June 27–29.

Back Hanna, Debus Marc, Fernandes Jorge M. (eds.) (2021), *The Politics of Legislative Debates*, Oxford: Oxford University Press.

Baden Christian, Pipal Christian, Schoonvelde Martijn, Velden Mariken van der (2022), *Three gaps in computational text analysis methods for social sciences: a research agenda*, "Communication Methods and Measures", vol. 16(1), pp. 1–18.

Beer Caroline (2009), *Democracy and gender equality*, "Studies in Comparative International Development", vol. 44, pp. 212–227.

Bischof Jonathan, Airoldi Edoardo M. (2012), *Summarizing topical content with word frequency and exclusivity*, [in:] John Langford, Joelle Pineau (eds.), *Proceedings of the 29th International Conference on Machine Learning*, Madison: Omnipress, pp. 9–16.

Blei David M., Ng Andrew Y., Jordan Michael I. (2003), *Latent Dirichlet Allocation*, "Journal of Machine Learning Research", vol. 3, pp. 993–1022.

Blum Avrim, Mitchell Tom (1998), *Combining labeled and unlabeled data with co-training*, [in:] Peter Barlett (ed.), *Proceedings of the 11th Annual Conference on Computational Learning Theory*, New York: ACM, pp. 92–100.

Boix Carles (2003), *Democracy and Redistribution*, Cambridge: Cambridge University Press.

Carlsen Hjalmar, Ralund Snorre (2022), *Computational grounded theory revisited: From computer-led to computer-assisted text analysis*, "Big Data and Society", vol. 9(1), pp. 1–16.

Ceka Besir, Magalhães Pedro C. (2020), *Do the rich and the poor have different conceptions of democracy? Socioeconomic status, inequality, and the political status quo*, "Comparative Politics", vol. 52(3), pp. 383–412.

Charmaz Kathy (2006), *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*, London: Sage Publications.

Coffé Hilde, Michels Ank (2014), *Education and Support for Representative, Direct and Stealth Democracy*, "Electoral Studies", vol. 35(1), pp. 1–11.

Coppedge Michael, Altman David, Fish Steven, Kroenig Matthew, McMann Kelly M., Gerring John, Bernhard Michael, Hicken Allen, Lindberg Staffan I. (2011), *Conceptualizing and measuring democracy: A new approach*, "Perspectives on Politics", vol. 9(2), pp. 247–267.

Cunningham Frank (2002), *Theories of Democracy: A Critical Introduction*, London: Routledge.

Dalton Richard J., Shin Doh Chull, Jou Willy (2007), *Popular Conceptions of the Meaning of Democracy: Democratic Understanding in Unlikely Places*, Irvine: Center for the Study of Democracy.

Deterding Nicole M., Waters Mary C. (2021), *Flexible coding of in-depth interviews: A twenty-first-century approach*, "Sociological Methods & Research", vol. 50(2), pp. 708–739.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina (2019), *Bert: Pre-training of deep bidirectional transformers for language understanding*, [in:] Jill Burstein, Christy Doran, Thamar Solorio (eds.), *ACL Anthology*, Minneapolis: ACL, pp. 4171–4186.

Dieng Adji B., Ruiz Francisco J.R., Blei David M. (2020), *Topic modeling in embedding spaces*, "Transactions of the Association for Computational Linguistics", vol. 8, pp. 439–453.

Dryzek John S. (1996), *Political inclusion and the dynamics of democratization*, "American Political Science Review", vol. 90(3), pp. 475–487.

Elster Jon, Offe Claus, Preuss Ulrich K. (1998), *Institutional Design in Post-Communist Societies: Rebuilding the Ship at Sea*, Cambridge: Cambridge University Press.

Erjavec Tomaž, Ogrodniczuk Maciej, Osenova Petya, Ljubešić Nikola, Simov Kiril, Pančur Andrej, Rudolf Michał, Kopp Matyáš, Barkarson Starkaður, Steingrímsson Steinþór, Çöltekin Çağrı, Does Jesse de, Depuydt Katrien, Agnoloni Tommaso, Venturi Giulia, Pérez María Calzada, Macedo Luciana D. de, Navarretta Costanza, Luxardo Giancarlo, Coole Matthew, Rayson Paul, Morkevičius Vaidas, Krilavičius Tomas, Darģis Roberts, Ring Orsolya, Heusden Ruben van, Marx Maarten, Fišer Darja (2023), *The ParlaMint corpora of parliamentary proceedings*, "Language Resources and Evaluation", vol. 57, pp. 415–448.

Ferrín Mónica, Kriesi Hanspeter (eds.) (2016), *How Europeans View and Evaluate Democracy*, Oxford: Oxford University Press.

Foster Ian, Ghani Rayid, Jarmin Ron S., Kreuter Frauke, Laneet Julia (eds.) (2021), *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*, Boca Raton: CRC Press.

Freedom House (2022), *Freedom in the World. The Global Expansion of Authoritarian Rule*, Washington: Freedom House.

Gillick Daniel, Presta Alessandro, Tomar Gaurav Singh (2018), *End-to-End Retrieval in Continuous Space*, https://arxiv.org/abs/1811.08008 [accessed: 21.02.2024].

Glaser Barney G., Strauss Anselm L. (1999), *The Discovery of Grounded Theory: Strategies for Qualitative Research*, New York: Aldine.

Guo Jiafeng, Cai Yinqiong, Fan Yixing, Sun Fei, Zhang Ruqing, Zhang Cheng (2022), *Semantic models for the first-stage retrieval: A comprehensive review*, "ACM Transactions on Information Systems (TOIS)", vol. 40(4), pp. 1–42.

Houle Christian (2015), *Ethnic inequality and the dismantling of democracy: A global analysis*, "World Politics", vol. 67(3), pp. 469–505.

Ilie Cornelia (2015), *Parliamentary discourse*, [in:] Karen Tracy (ed.), *The International Encyclopedia of Language and Social Interaction*, New Jersey: Wiley-Blackwell, pp. 1–15.

Jemielniak Dariusz (2020), *Thick Big Data: Doing Digital Social Sciences*, Oxford: Oxford University Press.

Katz Richard S. (2001), *Models of Democracy: Elite Attitudes and the Democratic Deficit in the European Union*, "European Union Politics", vol. 2(2), pp. 53–79.

Kherwa Pooja, Bansal Poonam (2019), *Topic modeling: a comprehensive review*, "EAI Endorsed Transactions on Scalable Information Systems", vol. 7(24), 159623.

Knutsen Carl Henrik, Wegmann Simone (2016), *Is democracy about redistribution?*, "Democratization", vol. 23(1), pp. 164–192.

Krippendorff Klaus (2018), *Content Analysis: An Introduction to its Methodology*, London: Sage.

Kwiatkowska Agnieszka (2017), *"Hańba w Sejmie" – zastosowanie modeli generatywnych do analizy debat parlamentarnych*, "Przegląd Socjologii Jakościowej", vol. XIII, no. 2, pp. 82–109.

Kwiatkowska Agnieszka, Grzybowska-Walecka Katarzyna (2024 forthcoming), *Polarized Democracy: Diverging Attitudes towards Democracy in Poland*, [in:] Katarzyna Grzybowska-Walecka, Simona Guerra, Fernando Casal Bértoa (eds.), *The Oxford Handbook of Polish Politics*, Oxford: Oxford University Press.

Kwiatkowska Agnieszka, Muliavka Viktoriia, Plisiecki Hubert (2023), *Hollowed or redefined? Changing visions of democracy in the political discourse of Law and Justice*, "Democratization", vol. 30(3), pp. 458–478.

Lehmann Pola, Franzmann Simon, Al-Gaddooa Denise, Burst Tobias, Ivanusch Christoph, Regel Sven, Riethmüller Felicia, Volkens Andrea, Weßels Bernhard, Zehnter Lisa (2024), *Manifesto Project Dataset (version 2024a),* Berlin: Wissenschaftszentrum Berlin für Sozialforschung, https://doi.org/10.25522/manifesto.mpds.2024a

Levitsky Steven, Ziblatt Daniel (2018), *How Democracies Die*, New York: Crown Publishing.

Lorenzini Jasmine, Kriesi Hanspeter, Makarov Peter, Wüest Bruno (2022), *Protest event analysis: Developing a semiautomated NLP approach*, "American Behavioral Scientist", vol. 66(5), pp. 555–577.

Lourenço Rui Pedro, Piotrowski Suzanne, Ingrams Alex (2017), *Open data driven public accountability*, "Transforming Government: People, Process and Policy", vol. 11(1), pp. 42–57.

Lyrio Maurício Vasconcellos Leão, Lunkes Rogério João, Castelló Taliani Emma (2018), *Thirty Years of Studies on Transparency, Accountability, and Corruption in the Public Sector: The State of the Art and Opportunities for Future Research*, "Public Integrity", vol. 20(5), pp. 512–533.

Markowski Radosław, Kwiatkowska Agnieszka (2018), *The Political Impact of the Global Economic Crisis in Poland: Delayed and Indirect Effects*, "Historical Social Research", vol. 43(4), pp. 250–273.

Meijer Albert Jacob (2003), *Transparent government: Parliamentary and legal accountability in an information age*, "Information Polity", vol. 8(1–2), pp. 67–78.

Meijer Harm Jan, Truong Joanne, Karimi Reza (2021), *Document Embedding for Scientific Articles: Efficacy of Word Embeddings vs TFIDF*, https://arxiv.org/abs/2107.05151 [accessed: 21.05.2024].

Mikolov Tomas, Kai Chen, Greg Corrado, Jeffrey Dean (2013a), *Efficient Estimation of Word Representations in Vector Space*, https://arxiv.org/abs/1301.3781 [accessed: 21.05.2024].

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, Dean Jeffrey (2013b), *Distributed Representations of Words and Phrases and their Compositionality*, https://arxiv.org/abs/1310.4546 [accessed: 21.05.2024].

Miłkowski Marcin (2022), *Morfologik software, version 2.1.6*, https://github.com/morfologik/morfologik-stemming/releases [accessed: 21.05.2024].

Munck Gerardo L., Verkuilen Jay (2002), *Conceptualizing and measuring democracy: Evaluating alternative indices*, "Comparative Political Studies", vol. 35(1), pp. 5–34.

Nelson Laura K. (2020), *Computational Grounded Theory: A Methodological Framework*, "Sociological Methods & Research", vol. 49(1), pp. 3–42.

Nicholls Tom, Culpepper Pepper D. (2021), *Computational identification of media frames: Strengths, weaknesses, and opportunities*, "Political Communication", vol. 38(1–2), pp. 159–181.

Nieuwelink Hessel, Dam Geert ten, Dekker Paul (2018), *Growing into politics? The development of adolescents' views on democracy over time*, "Politics", vol. 38(4), pp. 395–410.

Patton Michael Quinn (2014), *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, London: Sage Publications.

Plisiecki Hubert (2024), *Package retfidf. Document Retrieval for Social Sciences*, https://pypi.org/project/retfidf/ [accessed: 21.05.2024].

Plisiecki Hubert, Kwiatkowska Agnieszka (2022), *Finding democracy in big data: word-embedding-based document retrieval. Dataset*, https://osf.io/rk6pc [accessed: 21.05.2024].

Rauh Christian, Schwalbach Jan (2020), *The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies*, https://doi.org/10.7910/DVN/L4OAKN

Roberts Margaret E., Stewart Brandon M., Tingley Dustin, Lucas Christopher, Leder-Luis Jetson, Kushner Gadarian Shana, Albertson Bethany, Rand David G. (2014), *Structural topic models for open-ended survey responses*, "American Journal of Political Science", vol. 58(4), pp. 1064–1082.

Sack Benjamin C. (2017), *Regime change and the convergence of democratic value orientations through socialization. Evidence from reunited Germany*, "Democratization", vol. 24(3), pp. 444–462.

Saldaña Johnny (2021), *The Coding Manual for Qualitative Researchers*, London: Sage Publications.

Schmidt Craig W. (2019), *Improving a tf-idf weighted document vector embedding*, https://arxiv.org/abs/1902.09875 [accessed: 21.05.2024].

Schönhofen Peter (2009), *Identifying document topics using the Wikipedia category network*, "Web Intelligence and Agent Systems: An International Journal", vol. 7(2), pp. 195–207.

Schou Jannick, Hjelholt Morten (2018), *Digitalization and Public Sector Transformations*, Cham: Palgrave Macmillan.

Schwörer Jakob, Koß Michael (2023), *'Void' democrats? The populist notion of 'democracy' in action*, "Party Politics", https://doi.org/10.1177/13540688231200992

Scott Mike, Tribble Christopher (2006), *Textual Patterns: Key Words and Corpus Analysis in Language Education*, Philadelphia: John Benjamins.

Shaw Sylvia (2000), *Language, gender and floor apportionment in political debates*, "Discourse & Society", vol. 11(3), pp. 401–418.

Varieties of Democracy (2022), *Dataset v14 [Country-Year/Country-Date]. VoD Project*, https://doi.org/10.23696/mcwt-fr58

Voermans Wim, Napel Hans-Martien ten, Passchier Reijer (2015), *Combining efficiency and transparency in legislative processes*, "The Theory and Practice of Legislation", vol. 3(3), pp. 279–294.

Wang Di, Thint Marcus, Al-Rubaie Ahmad (2012), *Semi-supervised Latent Dirichlet Allocation and its application for document classification*, [in:] Li Yuefeng, Zhang Yanqing, Zhong Ning (eds), *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Los Alamitos: CPS, pp. 306–310.

Watanabe Kohei, Zhou Yuan (2022), *Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches*, "Social Science Computer Review", vol. 40(2), pp. 346–366.

Wodak Ruth, Krzyżanowski Michal (eds.) (2008), *Qualitative Discourse Analysis in the Social Sciences*, London: Palgrave MacMillan.

Wolf Thomas, Debut Lysandre, Sanh Victor, Chaumond Julien, Delangue Clement, Moi Anthony, Cistac Pierric, Rault Tim, Louf Remi, Funtowicz Morgan, Davison Joe, Shleifer Sam, Platen Patrick von, Ma Clara, Jernite Yacine, Plu Julien, Xu Canwen, Le Scao Teven, Gugger Sylvain, Drame Mariama, Lhoest Quentin, Rush Alexander (2020), *Transformers: State-of-the-art natural language processing*, [in:] Liu Qun, Schlangen David (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Minneapolis: ACL, pp. 38–45.

Xian Jasper, Teofili Tommaso, Pradeep Ronak, Lin Jimmy (2024), *Vector search with OpenAI embeddings: Lucene is all you need*, [in:] *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, New York: ACM, pp. 1090–1093.

Yamamoto Hironori (ed.) (2007), *Tools for Parliamentary Oversight: A Comparative Study of 88 National Parliaments*, Geneve: Inter-Parliamentary Union.

Young Iris Marion (2002), *Inclusion and Democracy*, Oxford: Oxford University Press.

Cytowanie

Hubert Plisiecki, Agnieszka Kwiatkowska (2024), *Discovering Representations of Democracy in Big Data: Purposive Semantic Sample Selection for Qualitative and Mixed-Methods Research*, „Przegląd Socjologii Jakościowej", t. XX, nr 4, s. 18–43, https://doi.org/10.18778/1733-8069.20.4.02

Odkrywanie reprezentacji demokracji w *Big Data*: semantyczny dobór celowy próby do badań jakościowych i mieszanych

Abstrakt: Wzrastająca liczba dużych, wielotematycznych korpusów tekstowych w naukach społecznych stwarza wyzwanie w doborze odpowiednich dokumentów do badań jakościowych i mieszanych. Tradycyjne metody doboru próby wymagają intensywnego kodowania manualnego lub uprzedniej wiedzy o zbiorze danych, podczas gdy metody nienadzorowane mogą dawać wyniki niespójne z kodowaniem opartym na teorii. Aby temu zaradzić, autorzy proponują semantyczny dobór celowy próby – podejście wykorzystujące przetwarzanie języka naturalnego z użyciem osadzeń dokumentów tworzonych przez średnią ważoną wektorów słów, z wagami określonymi współczynnikiem *tf-idf* (częstość terminu odwrotnie proporcjonalna do częstości dokumentu). Skuteczność podejścia zademonstrowano na przykładzie demokracji – złożonego tematu, trudnego do wydobycia z korpusów parlamentarnych. Proponowana metoda pozwala na niezawodny i efektywny dobór próby tekstów w dowolnej dziedzinie badań korzystającej z *Big Data*. Wkład autorów obejmuje walidację tego podejścia NLP dla nauk społecznych i humanistycznych oraz dostarczenie rzetelnego narzędzia dla badaczy, ułatwiającego pogłębioną analizę jakościową i eksplorację korpusów *Big Data* w ramach obliczeniowej teorii ugruntowanej.

Słowa kluczowe: dobór próby, dobór celowy, badania jakościowe, word embeddings, demokracja