Publizistik (2023) 68:325–354
https://doi.org/10.1007/s11616-023-00809-4

PUBLIZISTIK

AUFSATZ

The interactive Leipzig Corpus Miner: An extensible and adaptable text analysis tool for content analysis

Andreas Niekler · Christian Kahmann · Manuel Burghardt · Gerhard Heyer

Received: 4 January 2023 / Accepted: 11 July 2023 / Published online: 29 August 2023

Abstract We present the interactive Leipzig Corpus Miner (iLCM), which is the result of the development of an integrated research environment for the analysis of text data. The key features of iLCM compared to existing software tools for computer-assisted text analysis are its flexibility and scalability. The tool includes functions to offer commonly needed methods for automatic processing of text, such as preprocessing, standard text analysis, and visualization, which would be very time consuming without a ready-to-use software tool. To also provide more methodological flexibility, the iLCM is not tied to one specific class of research question, but can easily be extended to other applications. In this article, we will focus on the capabilities and the aspects of adaptability, extensibility, and data exchange with other tools from the field of empirical content analysis. We will present the features of the iLCM and showcase individual examples and a case study that demonstrates the practical use of the tool.

Keywords Automated Content Analysis · Text Mining · Natural Language Processing

✉ Dr. Andreas Niekler · Prof. Dr. Manuel Burghardt
Computational Humanities, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Germany
E-Mail: aniekler@informatik.uni-leipzig.de

Prof. Dr. Manuel Burghardt
E-Mail: burghardt@informatik.uni-leipzig.de

Dr. Christian Kahmann
Automatische Sprachverarbeitung, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Germany
E-Mail: kahmann@informatik.uni-leipzig.de

Prof. Dr. Gerhard Heyer
Sächsische Akademie der Wissenschaften zu Leipzig, Karl-Tauchnitz-Str. 1, 04107 Leipzig, Germany
E-Mail: heyer@saw-leipzig.de

Der interaktive Leipzig Corpus Miner: Ein erweiterbares und anpassungsfähiges Textanalysewerkzeug für die Inhaltsanalyse

Zusammenfassung Wir stellen den interaktiven Leipzig Corpus Miner (iLCM) vor, der das Ergebnis der Entwicklung einer integrierten Forschungsumgebung für die Analyse von Textdaten ist. Das Hauptmerkmal des iLCM, im Vergleich zu bestehenden Softwaretools für die computergestützte Textanalyse, ist seine Flexibilität und Skalierbarkeit. Das Tool enthält Funktionen, die häufig benötigte Methoden zur automatischen Verarbeitung von Texten anbieten, wie z. B. Vorverarbeitung, Standardtextanalyse und Visualisierung, die ohne ein Softwaretool sehr zeitaufwändig wären. Um auch eine größere methodische Flexibilität zu bieten, ist der iLCM nicht an eine bestimmte Klasse von Forschungsfragen gebunden, sondern kann leicht auf andere Anwendungen erweitert werden. In diesem Artikel werden wir uns auf die Eigenschaften und die Aspekte der Anpassbarkeit, der Erweiterbarkeit und des Datenaustauschs mit anderen Instrumenten aus dem Bereich der empirischen Inhaltsanalyse konzentrieren. Wir stellen die Funktionen des iLCM vor und zeigen ein Beispiel, das den praktischen Einsatz des Tools demonstriert.

Schlüsselwörter Automatisierte Inhaltsanalyse · Text Mining · Automatische Sprachverarbeitung

1 Introduction

Text is a focal point of study in content-oriented social sciences research and communication studies, serving as one of its primary research subjects. In addition, researchers commonly employ automatic content analysis as a research method to explore extensive volumes of textual data. From a historical perspective, the field of computational social science (CSS) has emerged over the past decade, taking advantage of the vast amounts of digital data now accessible on the internet, which includes diverse sources like newspapers, parliamentary protocols, social media platforms, administrative records, and historical archives. The term itself was coined by Lazer et al. in 2009, who identified CSS as a broad field with a focus on network analysis and the aim of better understanding both the structure and content of relationships. More recently, in a review of computational social science and sociology, Edelmann et al. (2020, p. 64) demonstrated that we can observe a rapid growth of new techniques and tools since 2010 that help to analyze these large, complex datasets; in particular, various forms of automated text analysis for the ever-increasing amount of textual data (for example, see the text books of Ignatow and Mihalcea 2016 and Macanovic 2022; for an example of such a study, see Wiedemann 2016). Similarly, an increasing number of resources have been developed to support training and education in these emerging methodological paradigms (for instance, see the Language Technology and Data Analysis Laboratory at the University of Queensland, https://ladal.edu.au; Wiedemann and Niekler 2017).

While more and more researchers in the (digital) humanities and (computational) social sciences continue to embrace programming skills, it is worth noting that

the development of sophisticated research software remains a complex undertaking, often necessitating the expertise of trained software engineers. As a consequence, applied research mostly relies on ready-to-use research software that is typically designed for a specific task or method, such as creating concordances and calculating word frequencies (for example, https://www.laurenceanthony.net/software/antconc/; on which, see Anthony 2005) or for creating topic models (for example, https://dariah-de.github.io/TopicsExplorer/; on which, see Simmler et al. 2019). While such tools are typically easy to use, they also impose limitations on researchers, confining their study designs to the predefined boundaries of the software. Consequently, there is limited flexibility to adapt the methods to accommodate more intricate research design.

As an alternative to such immediately applicable yet relatively static and inflexible tools, one may instead advocate for the use of more versatile software packages that implement fundamental text mining methods. As Grimmer and Stewart (2013, pp. 267–297) noted, clustering methods and supervised or unsupervised methods for text classification, often based on prior human hand coding of documents into a predetermined set of categories, are key elements of computer-supported systematic analysis of large-scale text collections. Most of the methodological requirements Grimmer mentioned are in fact implemented in tools and open source software packages for Natural Language Processing (NLP), such as StanfordNLP (see Manning et al. 2014), OpenNLP (http://opennlp.apache.org/), NLTK (see Bird et al. 2009), Gensim (see Řehůřek and Sojka 2010), SpaCy (see Honnibal et al. 2022), or Quanteda (see Benoit et al. 2018), and have been around for the last decade.

However, employing these software packages for intricate workflows can pose technical challenges, demanding a profound understanding of the frameworks and their associated application programming interfaces (APIs). The level of expertise required might act as a deterrent for many researchers. In order to make such technically demanding software frameworks more accessible, several attempts have been made to integrate rudimentary NLP pipelines into ready-to-use tools, such as Clarin Weblicht (see Heyer and Böhlke 2021; Hinrichs et al. 2010), Triple (see Dumouchel et al. 2020), or Textgrid (see Neuroth et al. 2015). Nevertheless, this approach leads to similar limitations as those of the static tools mentioned earlier, as researchers can only recombine a predetermined set of processes within the tool, without the ability to fundamentally modify or expand upon them.

It should also be noted that there are numerous tools available for qualitative data analysis (QDA), such as VERBI software's MAXQDA (https://maxqda.com/), ATLAS.ti (https://atlasti.com), or NVivo (see Richards 2021). These tools are specifically devised to support QDA research design, involving the manual annotation of texts based on a provided codebook. These software solutions generally lack the flexibility to accommodate research designs that deviate from this traditional approach, such as incorporating automatic machine learning methods for text classification during the coding process.

Due to the growing need and challenges identified, we have designed the interactive Leipzig Corpus Miner (iLCM; also see Niekler et al. 2014, 2018), which is the result of the development of an integrated research environment for the analysis of text data (https://ilcm.informatik.uni-leipzig.de/). The key features of the iLCM

compared to existing software tools for computer-assisted text analysis are its flexibility and scalability. Most importantly, the tool's functionality offers commonly needed methods for automatic processing of text—such as preprocessing, standard text analysis, and visualization—which would be time consuming without a ready-to-use software tool. In order to also provide more methodological flexibility, the iLCM is not tied to one specific class of research question, but can easily be ported to other applications. Users can initially explore the tool's functions through an easy-to-use graphical user interface (GUI) and customize or expand specific features as needed.

The iLCM's extensibility is made possible because it is built entirely using the open-source environment R (https://www.r-project.org/). This means that all the tool's functions are based on modular R scripts. Because these are "hidden" behind an RShiny-based (https://shiny.rstudio.com/) GUI, researchers can choose whether or not they wish to modify the predefined scripts, doing so based on their needs or abilities. To facilitate this customization, the tool provides an environment within its GUI where the integrated R scripts can be edited and saved as custom scripts. When initiating analysis, researchers can then utilize these custom scripts instead of the standard ones. This customizability and extensibility also allow researchers to integrate a wide range of relevant text mining methods that are already available via R implementations into the tool. As a result, the functionality of the software offers more than existing standalone tools, inasmuch as it brings a range of functionality together through the one GUI.

The range of functions offered by the iLCM includes, among others, retrieval and management of document collections, the analysis of word frequency, time series analysis, topic models, and the automatic coding and annotation of categories, or supervised text classification as a "Software as a Service" architecture (SaaS). Its built-in ability to produce custom scripts and to export results and script-based adaptations of the available analyses circumvents some restrictions of other tools for text-oriented analysis methodologies. In short, the iLCM research environment addresses 1) the requirements for quantitative analysis of large qualitative data using text mining methods and 2) the requirements for reproducibility, intersubjectivity, and validity of data-driven research design in the social sciences.

In this article, we will focus on the iLCM's capabilities and adaptability, extensibility, and data exchange with other tools from the field of empirical content analysis. We will first present the features of the iLCM and showcase individual examples. In addition to providing an overview of the tool's functionality, in Section 3, we will showcase the practical use of the iLCM in a real research project within the field of communication studies. This case study will exemplify the utilization of iLCM across all stages of the research process, offering a comprehensive illustration of its capabilities. The text is thus intended to help the reader learn about the methods implemented and to generate an understanding of how the different functions can contribute to different research paradigms.

2 Overview of the interactive Leipzig Corpus Miner (iLCM)

The iLCM provides a variety of different functionalities that are useful for dealing with large text corpora. First and foremost, it serves as a text mining infrastructure specifically tailored to content-based research tasks. It is based on the Leipzig Corpus Miner, which was developed as part of an interdisciplinary project titled "Postdemokratie und Neoliberalismus" (ePol; Wiedemann et al. 2013). The aim of the ePol project was to analyze over 3.5 million news items from 60 years of German newspaper history. The iLCM, on the other hand, has a broader focus. It was initially based on the idea of combining quantitative (e.g., exploratory search or automatic classification) and qualitative approaches (e.g., through manual annotation of textual documents on the basis of a codebook) in a single tool to support extensive mixed method approaches. To this end, a number of different options have been implemented for the analysis of textual data.

In this section, we describe the tool's capabilities in detail and contextualize their usage for different research tasks in content analysis. We describe the main functions of the iLCM, aligning them with a typical research workflow in social sciences and communication studies (see Fig. 1). The workflow graphic illustrates that the iLCM comes with a wide variety of functions and methods to support researchers throughout the research process. We will briefly summarize these functions and, when appropriate, provide methodological reflections and usage examples.

2.1 Installation

As the iLCM consists of a number of different modules, its setup can be challenging for different operating systems. To ensure an easy set up, we decided to utilize virtualization. This means the iLCM comes as an image, which was previously defined by the developer as a self-contained environment that includes all the necessary libraries and dependencies. The end user is therefore not required to manually install all the necessary software packages, but merely needs to install the software to execute such images, independent of the used operating systems.

Specifically, we utilized the Docker framework to develop an image that can be downloaded and launched with a single command (https://hub.docker.com/r/ckahmann/ilcm/tags). Although setting up the iLCM on desktop machines in this manner is convenient, it is important to acknowledge that not all machines are equally suitable, and the availability of computing resources significantly impacts the handling of large and complex data. Consequently, for projects reliant on extensive datasets, we highly recommend running the iLCM on suitable server environments.

2.2 Data import

The iLCM's flexible import and export interface allows users to enter data in structured (CSV, Excel, Rotterdam Exchange Format Initiative [REFI]) as well as unstructured (PDF, DOCX, TXT) formats. No single standard format is required. Instead, the tool interactively maps the existing data structure, including the given

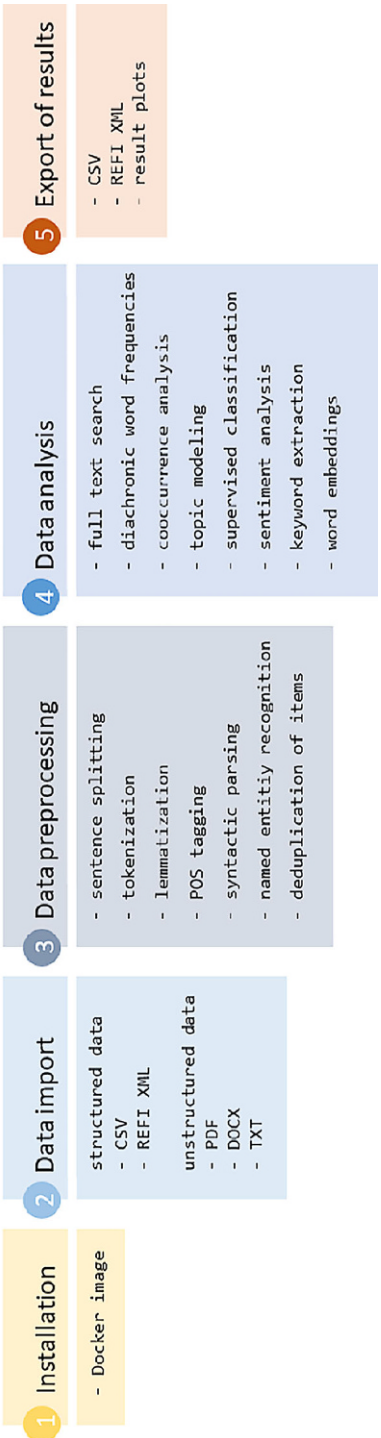


Fig. 1 Stages of a typical research workflow in social sciences and communication studies aligned with the functionality that the iLCM provides for each of these stages

metadata, to the internally used data format. An interface is available in the tool to assign which information is mapped to the internal data fields of the iLCM.

During the import process, it is essential to specify at least the date and title of the document. However, the iLCM allows for the inclusion of up to nine additional metadata fields for more comprehensive document organization. In cases where the date or title is unavailable, the iLCM provides options to set these automatically or with the assistance of an R script. This ensures a smooth and efficient import process for various types of documents.

2.3 Data processing

Within the iLCM environment, there are various data processing mechanisms available, encompassing linguistic preprocessing as well as a deduplication process for the text data. The linguistic preprocessing of the text data is carried out automatically with the *SpaCy* library when importing new data into the tool. This includes *sentence segmentation, tokenization*, and *lemmatization* as well as *part of speech* (POS) *tagging, syntactic parsing*, and *named entity recognition* (NER). Large parts of the implemented functions of the iLCM rely on such pre-processed linguistic information. Besides the need to be able to perform sentence- or word-level analysis, it is, for example, beneficial to be able to provide meta information (POS, named entities, grammatical categories) for each word during the annotation phase. The results of the preprocessing are stored in the database so that they are available for later analysis without having to process the text data repeatedly.

SpaCy enables the use of pre-trained models for various languages. By default, the iLCM includes models for English and German. However, if additional language models are required, users can easily add these with just a few clicks. The language options interface in the iLCM displays the models that are currently installed and provides a straightforward installation process for additional language models.

In the context of content analysis and the application of text-mining methods, deduplication plays a crucial role in ensuring the reliability and significance of results by establishing a duplicate-free corpus. As Benko has argued (2013, p. 27), the increased availability of web corpora, particularly those compiled partially or fully through automated methods, has underscored the necessity of document deduplication in content analysis. Duplicates in text can greatly distort procedures, such as co-occurrence analysis, frequency analysis, and topic modeling; the presence of duplicate words can have a substantial impact on accuracy and validity. Depending on the extent of duplicate occurrences in the text, the entire analysis of the corpus using automatic language processing methods may become unreliable or even obsolete without proper document deduplication.

2.4 Data analysis

Upon importing and processing the text data, the iLCM enables the application of various text mining and machine learning methods for analysis. The following sections provide a concise overview of the different functions.

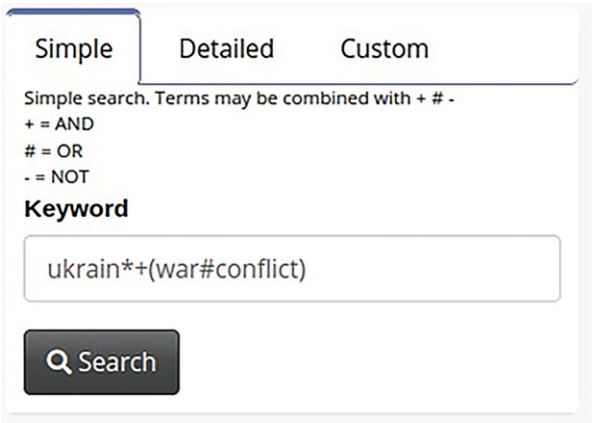


Fig. 2 Search interface of the iLCM, in which complex search queries can be built by using operators such as "*" (Wildcards), "AND", "OR", and "NOT"

2.4.1 Full text search

After importing the corpus into the tool, the iLCM offers functions for document search and the creation of sub-corpora known as collections. Often, only specific portions of a corpus are relevant to a given research question. Thus, to facilitate filtering, the iLCM allows for real-time search queries that combine complex keyword searches with existing metadata conditions. This enables the creation of customized searches based on specific criteria. Additionally, the iLCM provides options for sampling documents into collections, which is particularly useful when working with extensive amounts of text data. For instance, if the analysis of a given corpus were to focus only on those texts that are somehow related to the war in Ukraine, one can narrow down the corpus using specific search terms, as shown in Fig. 2.

2.4.2 Diachronic word frequencies

One popular approach to analyzing text data is studying diachronic word frequencies (see, for example, Michel et al. 2011). This involves tracking the frequency curve of specific words or word groups over time and comparing them. This perspective allows researchers to identify the most frequent words within different temporal aggregation ranges. Frequency data can be measured at the word and document levels. Researchers can analyze absolute frequencies, based on time points, or normalized frequencies, based on document quantities at different time points.

The following brief example further illustrates the iLCM's capabilities with regards to this approach. Here, diachronic word frequency is applied to analyze the mentions of political parties over time in texts from the German-language newspaper *taz*. This was part of an investigation into the correlation between the number of mentions and the respective parties' poll results. Figure 3 displays the monthly number of mentions for four selected party names. This counting approach can be expanded by utilizing dictionaries, which enable the inclusion of additional synonyms or the names of top politicians associated with the parties. By doing so,

Fig. 3 The visualization interface for diachronic frequency analysis in the iLCM. Here, the frequencies of appearance of various German political parties in 2017 are presented by month. The data was retrieved from the archive of the German-language newspaper taz. The designations cdu, spd, afd and csu are abbreviations for German parties that are examined here

the analysis goes beyond counting official party names and encompasses a broader understanding of party mentions in the text data.

2.4.3 Co-occurrence analysis

Another popular way to explore large text corpora involves the analysis of significant co-occurrences of words, which can be calculated with various statistical measures in the iLCM (see Biemann et al. 2022 for more details on significance measures). Results of co-occurrence analysis are displayed both by means of a *Keyword in Context* view of the words as well as a network visualization, making it easy to discover patterns of co-occurring concepts in a corpus. As an advanced co-occurrence feature, the iLCM also provides a measure called context volatility (see, for example, Heyer et al. 2009), which takes into account how much the context of a word (i.e., its co-occurrences) change over the course of time (= volatility).

The example in Fig. 4 shows a co-occurrence network at sentence level for the word "water". The corpus includes texts from the Nationally Determined Contributions (NDCs), which are the heart of the Paris Agreement and "embody efforts by each country to reduce national emissions and adapt to the impacts of climate change" (United Nations Framework Convention on Climate Change 2022). The representation as a network makes it possible to quickly determine the most important related words of the keyword under investigation.

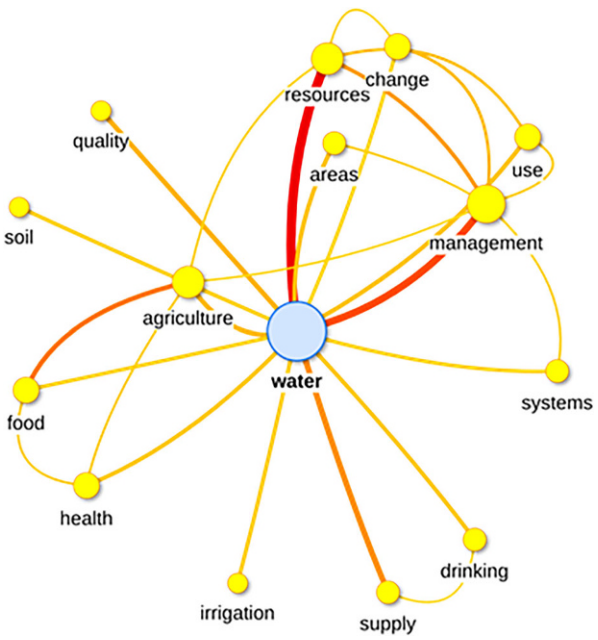


Fig. 4 Network visualization for the word "water" and its statistically significant co-occurrences in the NDCs corpus

2.4.4 Topic modeling

Topic models offer an unsupervised approach for clustering documents. The method is based on the *Latent Dirichlet Allocation* (LDA) algorithm, first presented by Blei et al. (2003). LDA models the word compositions of documents and organizes them into coherent groups (= topics) based on word usage. In essence, the model assigns a probability distribution to each document indicating its likelihood of belonging to different topics and assigns a probability distribution to each inferred topic indicating its usage of words from the entire vocabulary. Chen et al. (2023) offer a systematic review of what can (and cannot) be done with the topic modeling method in communication studies. In addition to LDA topic modeling, the iLCM also offers *Dynamic Topic Modeling* (cf. Blei and Lafferty 2006) and uses *Structured Topic Models* (STM; cf. Roberts et al. 2016).

The iLCM offers multiple approaches for evaluating and assessing the quality of topic models which is necessary for a valid application. Maier et al. (2018) give an overview of a valid methodological approach in automatic content analysis. These include measures such as topic coherence and topic intrusion, which provide insights into the coherence and relevance of topics. Additionally, the tool can check the topic reliability, allowing researchers to examine the consistency and stability of the identified topics.

The iLCM also provides capabilities for evaluating correlations between metadata and topic distributions. It allows for the mapping of diachronic trends in topic distributions at specific time points. For qualitative analysis, researchers can select the most relevant documents for a chosen topic, which are then displayed with key words highlighted in the corresponding color. This enables quick identification and verification of particularly relevant text passages. This approach enhances the interpretation of topics beyond simple word lists and helps to identify potential systematic errors in the data or the modeling process. To facilitate interpretation, the iLCM includes a labeling tool that allows researchers to assign uniform names to topics after an interpretation step has been conducted.

An application scenario for topic modeling could identify the main thematic discussion items in a set of texts and evaluate their distribution over the corpus. In the excerpt shown in Fig. 5, national climate strategies were evaluated. The text was downloaded from different online sources and was provided by the research project TRANSNORMS (see www.transnorms.eu). Topics such as “funding”, “renewable energy”, or “waste management” emerged and can be further analyzed with respect to metadata or accompanying information in the text data.

2.4.5 Supervised classification

One of the fundamental requirements in quantitative content analysis is the process of coding a measurable variable into identifiable categories within the texts being analyzed (see Früh 2001; p. 80; Krippendorff 2018, pp. 155–161). Thus, in addition to unsupervised clustering through topic modeling, the iLCM also offers integrated procedures for supervised classification—a machine learning technique where documents are assigned predefined labels based on a training dataset. Researchers can

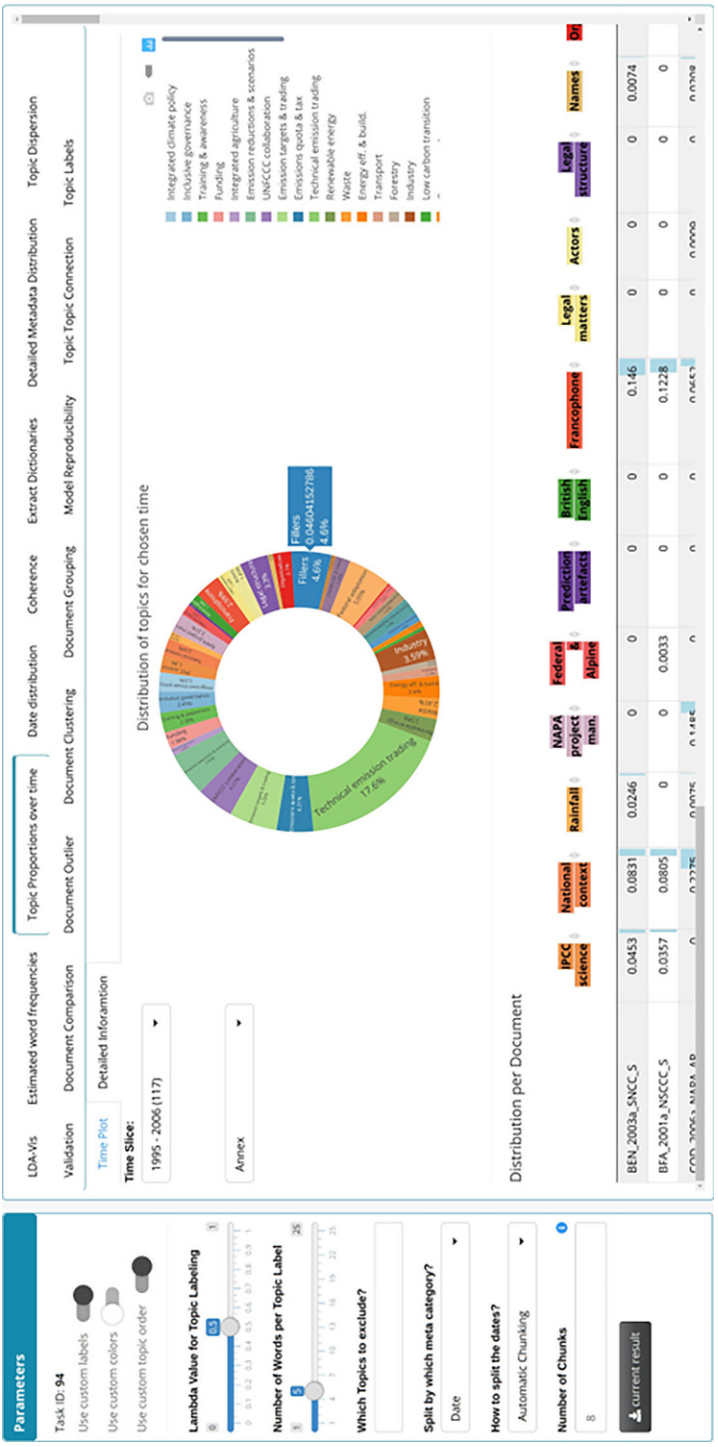


Fig. 5 The image shows the interface for the evaluation of topic model results. The distribution of topic importance for a specified period is shown. The topics were previously labeled using the available labeling tool

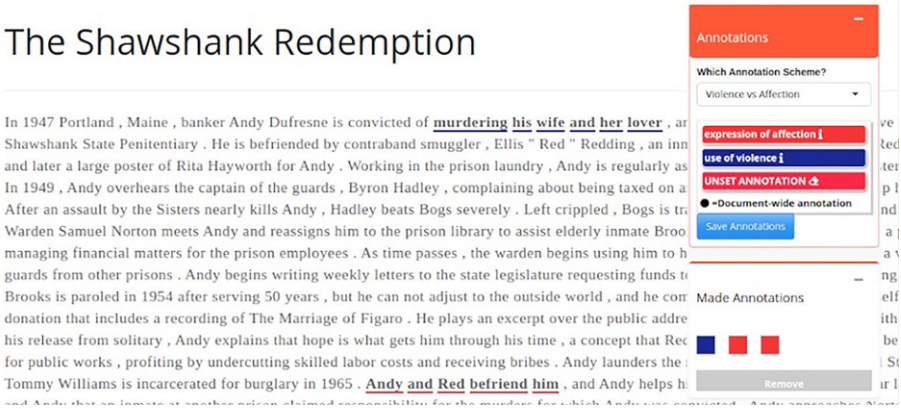


Fig. 6 An illustration of the document view in the iLCM, where textual evidence can be annotated according to the categories (affection, violence) of a selected codebook

create codebooks within the iLCM's GUI, adhering to the requirements of content analysis procedures. Using the annotation interface, documents can be annotated based on these codebooks. These annotations serve as a training dataset for building a classifier.

Additionally, the iLCM supports the initiation of an Active Learning (AL) process (see Settles 2012; Schröder et al. 2022) based on an initial training set or a dictionary-based search. With the help of a classifier, new examples for AL can be generated. This approach suggests potential instances of texts for different codebook categories to the user automatically. This significantly reduces time and effort in comparison to manual qualitative annotation. AL facilitates the efficient creation of a sufficiently large training set, enabling the classifier to be applied accurately to entire document sets. The results of this classification can be further examined quantitatively, exported to other tools, or utilized in subsequent analysis, such as co-occurrence calculations based on the classification examples.

The following example showcases the application of the iLCM's supervised classification functions, examining the changing relationship between the portrayal of affection and violence in movies over the past four decades. The analysis draws on a dataset consisting of short descriptions of movies from the past 40 years (derived from Kaggle 2022). A set of labels, according to a given codebook, with the two categories *affection* and *violence* is used (Fig. 6) to annotate the data and AL was employed to quickly expand the training dataset. Finally, the supervised classifier was applied to the entire set of movie descriptions, the result of which can be seen in Fig. 7. Based on the movie descriptions, we can observe that the concepts of *affection* and *violence* both show a rising trend from the year 2000. Furthermore, we can observe a larger proportion of content reflecting the concept of *affection*.

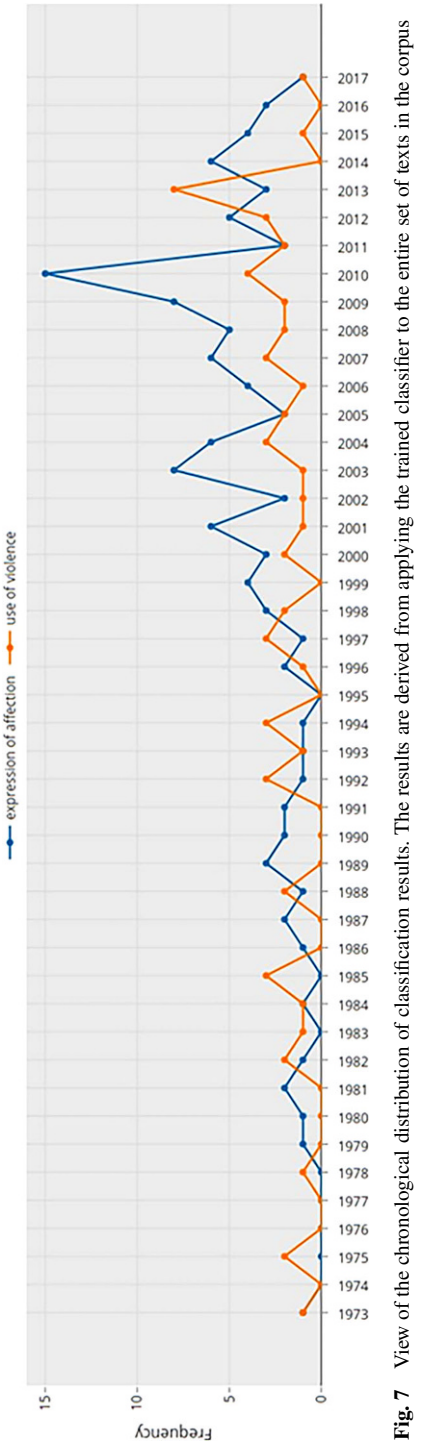


Fig. 7 View of the chronological distribution of classification results. The results are derived from applying the trained classifier to the entire set of texts in the corpus

2.4.6 Sentiment analysis

Sentiment analysis, which involves quantifying the emotional tone expressed in a text, plays a crucial role in understanding the emotionality of media content. By analyzing sentiment, we can explain various media effects and gain insights into how emotional aspects shape the presentation of topics, political campaigns, and historical events. Recognizing the influence of emotions in media content is essential for comprehending their impact and implications in shaping audience perceptions and societal narratives (see Döveling and Konijn 2021, pp. 48–66; Kühne et al. 2021, p. 128; Nabi 2019, pp 163–178).

Sentiment dictionaries (on which, see Khoo and Johnkhan 2018; Ribeiro et al. 2016) are curated collections of words or phrases along with their associated sentiment scores. Such dictionaries serve as a reference for sentiment analysis tasks, allowing the classification of text based on the presence of positive, negative, or neutral words, and providing a basis for quantifying the sentiment expressed in a given text. By default, sentiment dictionaries for English and German are available in the iLCM; however, these can also be extended or supplemented as desired.

After extracting the document's sentiments based on a dictionary, results can be evaluated according to the metadata. Figure 8 illustrates the iLCM's sentiment analysis function with an example of a textual description of a movie (derived from Kaggle 2022). We chose the movie *Seven*, since its emotional content draws a threatening scenario. This example also nicely shows that sentiment analysis does not always have to be about value judgements, but simply looks at emotional components, whether value judgements or expressed sentiment.

2.4.7 Keyword extraction

By extracting keywords (i.e., terms that are characteristic for a specific text), we can efficiently obtain a concise summary of essential information within a document collection. Keyword extraction methods use statistical characteristics of word distributions to assign weights to words based on their statistical significance. In the iLCM, standard keyword extraction procedures such as *RAKE* or *Textrank* are available to users (on which, see Ganiger and Rajashekharaiah 2018).

Figure 9 showcases keyword extraction from a news corpus of the British daily newspaper *The Guardian* using the RAKE algorithm. This analysis provides a swift understanding of the primary focus of the reporting, which revolves around the handling and impact of the Covid-19 pandemic. In this analysis, keywords have been selected to be bigrams (that is, two-word phrases such as "new cases", "press conference", "first dose", etc.).

2.4.8 Word embeddings

Since the release of *Word2Vec* by Mikolov et al. (2013), word embedding techniques have significantly improved all kinds of NLP tasks. Word embeddings are dense vector representations of words in a high-dimensional space, where words with similar

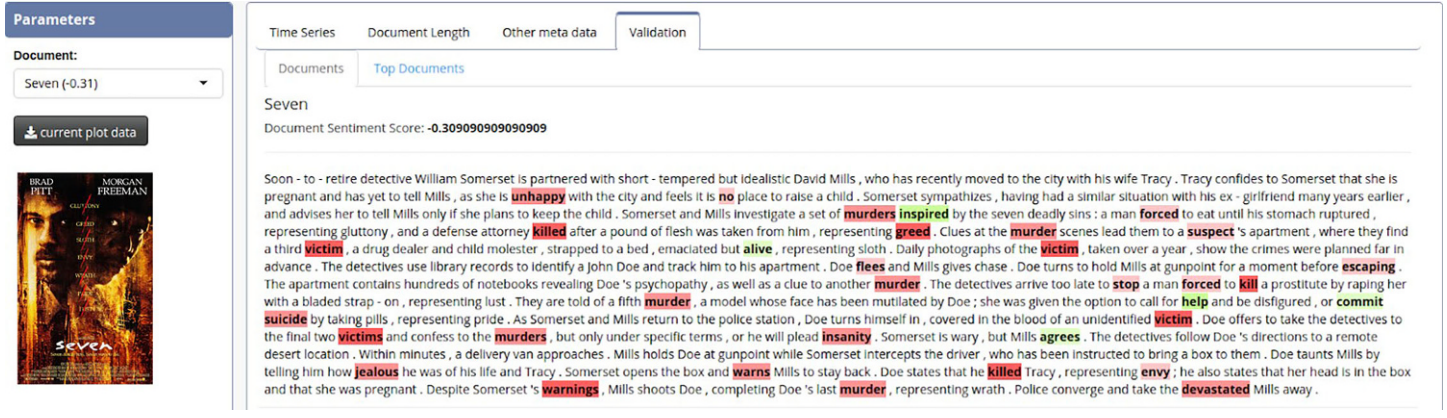
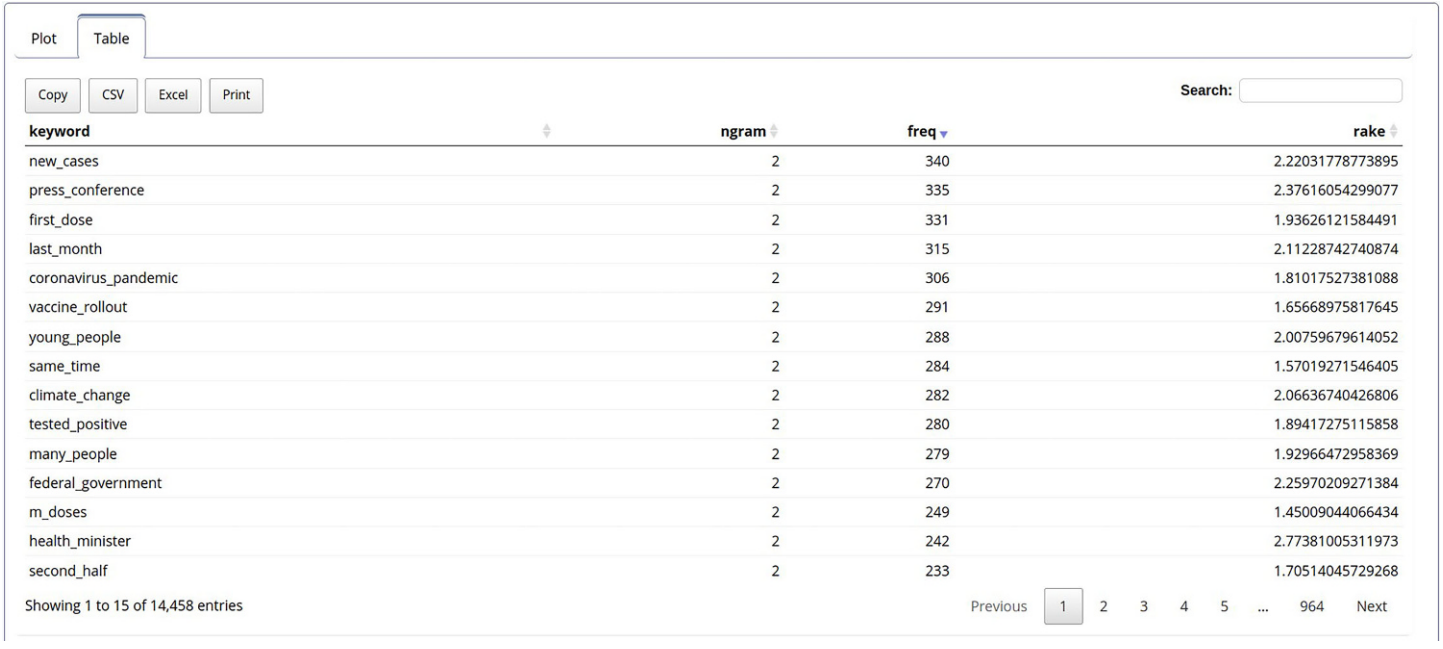


Fig. 8 A detailed view of the sentiment analysis results, with red-highlighted words indicating sentiments with negative connotations and green-highlighted words representing those with positive connotations



Plot | Table

Copy | CSV | Excel | Print

Search:

keyword	ngram	freq	rake
new_cases	2	340	2.22031778773895
press_conference	2	335	2.37616054299077
first_dose	2	331	1.93626121584491
last_month	2	315	2.11228742740874
coronavirus_pandemic	2	306	1.81017527381088
vaccine_rollout	2	291	1.65668975817645
young_people	2	288	2.00759679614052
same_time	2	284	1.57019271546405
climate_change	2	282	2.06636740426806
tested_positive	2	280	1.89417275115858
many_people	2	279	1.92966472958369
federal_government	2	270	2.25970209271384
m_doses	2	249	1.45009044066434
health_minister	2	242	2.77381005311973
second_half	2	233	1.70514045729268

Showing 1 to 15 of 14,458 entries

Previous | 1 | 2 | 3 | 4 | 5 | ... | 964 | Next

Fig. 9 Results for a 2gram keyword analysis of Guardian newspaper articles that mainly report on the Covid-19 pandemic

meanings are positioned closer to each other. They capture semantic relationships and contextual information, facilitating various NLP tasks.

The iLCM incorporates the conventional Word2Vec methodology and additionally enables users to import alternative word embedding models. These embedding models' outcomes are then projected from a multi-dimensional vector space into a two-dimensional visualization, making it easily interpretable for researchers. The iLCM provides two well-known approaches for achieving this dimensionality-reduction process: *Principal Component Analysis* (PCA; see Jolliffe and Cadima 2016) and *t-Distributed Stochastic Neighbor Embedding* (t-SNE; see Maaten and Hinton 2008).

For the example shown in Fig. 10, we trained a Word2Vec model on the national climate strategies of countries grouped by membership in different negotiating groups (cf. United Nations Framework Convention on Climate Change 2022). The figure shows a comparison table with the most similar words to the initial word "forestry" according to different groups. Preliminary indications of the focus of the different negotiation groups can thereby be extracted. For example, we see that the focus of the OPEC or Umbrella Groups seems to be more industrial, while in the group of EU countries there is a more general focus.

2.5 Export of results

In some cases, it may not be feasible to conduct a research project exclusively within the iLCM. Additionally, there are other software solutions that excel in implementing specific methods, and it would be unfortunate to overlook their potential. In such situations, the extensive input and export interfaces of the iLCM allow for the extraction of results or intermediate findings, which can then be seamlessly processed in other software environments according to specific needs and requirements.

Various results such as frequency, time series, topic model outputs, or classified texts can be directly exported from the iLCM into CSV format. The iLCM also provides the capability to export data in the REFI format, enabling further evaluations and analysis to be conducted using traditional QDA software solutions. This integration allows researchers to leverage the strengths of both the iLCM and QDA software for a mixed methods approach to data exploration.

3 Case study: investigating speech acts in political communication

The following case study illustrates some of the previously introduced functions of the iLCM with a real-world research scenario. Specifically, we will use the iLCM to explore the utilization of speech acts in the realm of international political communication.

When we engage in language, our intention often extends beyond simply transmitting information, as we aim to achieve a communicative objective and fulfill a social purpose, such as making a commitment or delivering a warning. Rather than merely expressing words, we actively use language to perform actions and achieve specific purposes (Austin 1962, pp. 6–7). Searle (1976) expanded upon this

EU	Umbrella_group	EIG	LDC	LLDC	Cartagena_Dialogue	AOSIS	CfRN	LMDC	BASIC	OPEC
(0.3) change	(0.36) industry	(0.18) landuse	(0.24) landuse	(0.3) agriculture	(0.3) landuse	(0.29) agriculture	(0.29) agriculture	(0.27) agriculture	(0.35) agriculture	(0.45) plantations
(0.38) landuse	(0.36) change	(0.26) agriculture	(0.24) agriculture	(0.32) land_use	(0.31) agriculture	(0.37) landuse	(0.33) meteorology	(0.3) landuse	(0.44) water	(0.49) restoring
(0.52) solvent	(0.39) landuse	(0.27) power	(0.29) land_use	(0.35) landuse	(0.37) aquaculture	(0.39) land_use	(0.34) hunting	(0.35) silviculture	(0.44) intellectual	(0.51) aquaculture
(0.52) land_use	(0.42) cement	(0.3) change	(0.37) land_uses	(0.38) land_uses	(0.42) grazing	(0.42) fisheries	(0.35) finally	(0.36) livestock	(0.45) ecosystems	(0.52) techniques
(0.55) agriculture	(0.42) storage	(0.31) waste	(0.39) husbandry	(0.39) husbandry	(0.43) incl	(0.44) change	(0.39) tourism	(0.38) agricultural	(0.46) protected	(0.52) nurseries
(0.57) manufacture	(0.43) land_use	(0.34) generation	(0.39) finally	(0.41) sector	(0.44) change	(0.45) sector	(0.39) fishing	(0.42) land_use	(0.46) supporting	(0.53) watershed
(0.6) removals	(0.43) product	(0.4) urban	(0.4) subsectors	(0.41) fishery	(0.47) livestock	(0.47) waste	(0.4) suburban	(0.43) poultry	(0.48) region	(0.55) harvesting

Fig. 10 Illustration of the result view for Word2Vec models. The most similar words to the given initial word "forestry" per negotiation group are shown

conception and introduced the broader term "speech acts". In the realm of political communication, speech acts play a pivotal role, as they serve as powerful tools employed by politicians to enact specific actions and shape the perceptions and behaviors of their audience. In the following case study, we will focus on two speech act types, namely *expressives*, where the speaker conveys personal thoughts and emotions, and *declaratives*, where the speaker issues orders or instructions to the recipients.

3.1 Data import and data processing

The corpus used for this case study describes over 8000 speeches held at the general debate of the United Nations General Assembly (UNGA) between 1970 and 2020 (https://doi.org/10.7910/DVN/0TJX8Y; also see Baturo et al. 2017). The speeches are all in English, and in addition to the actual text, are also tagged with metadata such as speaker, country, year, and session. The data is available as CSV file, where each line describes a single speech.

To import the corpus into the iLCM, the tool offers an input interface. Once the upload is complete, we interactively map the available data to match the metadata standard used in the iLCM (see Fig. 11). The iLCM's interactive mapping process eliminates the need for explicitly defined data formats, which is a common constraint in many other systems. With the iLCM, every imported text undergoes automatic preprocessing, as described in Sect. 2.3.1 (sentence splitting, tokenization, lemmatization, POS tagging, syntactic parsing, NER).

3.2 Data analysis

3.2.1 Full text search

As we are interested in the analysis of speech acts in political communication, the first task is to check for the general availability of speech acts in the corpus. This can easily be achieved by looking up specific performative verbs, which, as Searle (1976, p. 16) has argued regularly co-occur with speech acts. To gain a first overview of the corpus and potential speech acts, we use the full-text search, looking for the verbs "will" and "support", which are often associated with declarative speech acts. With a total of 7625 sentences containing at least one of the targeted verbs, we can now leverage the iLCM to conduct a close reading of selected sentences, ensuring that they accurately represent the desired speech act types. Two example sentences found by the iLCM are shown here:

> We **support** the peoples of Mozambique and Angola in their struggle to defend national independence against interference and aggression by imperialists and their reactionary lackeys. Laos, 1977
> We **will** continue to combat illicit trade and the spread of small arms. Norway, 2000

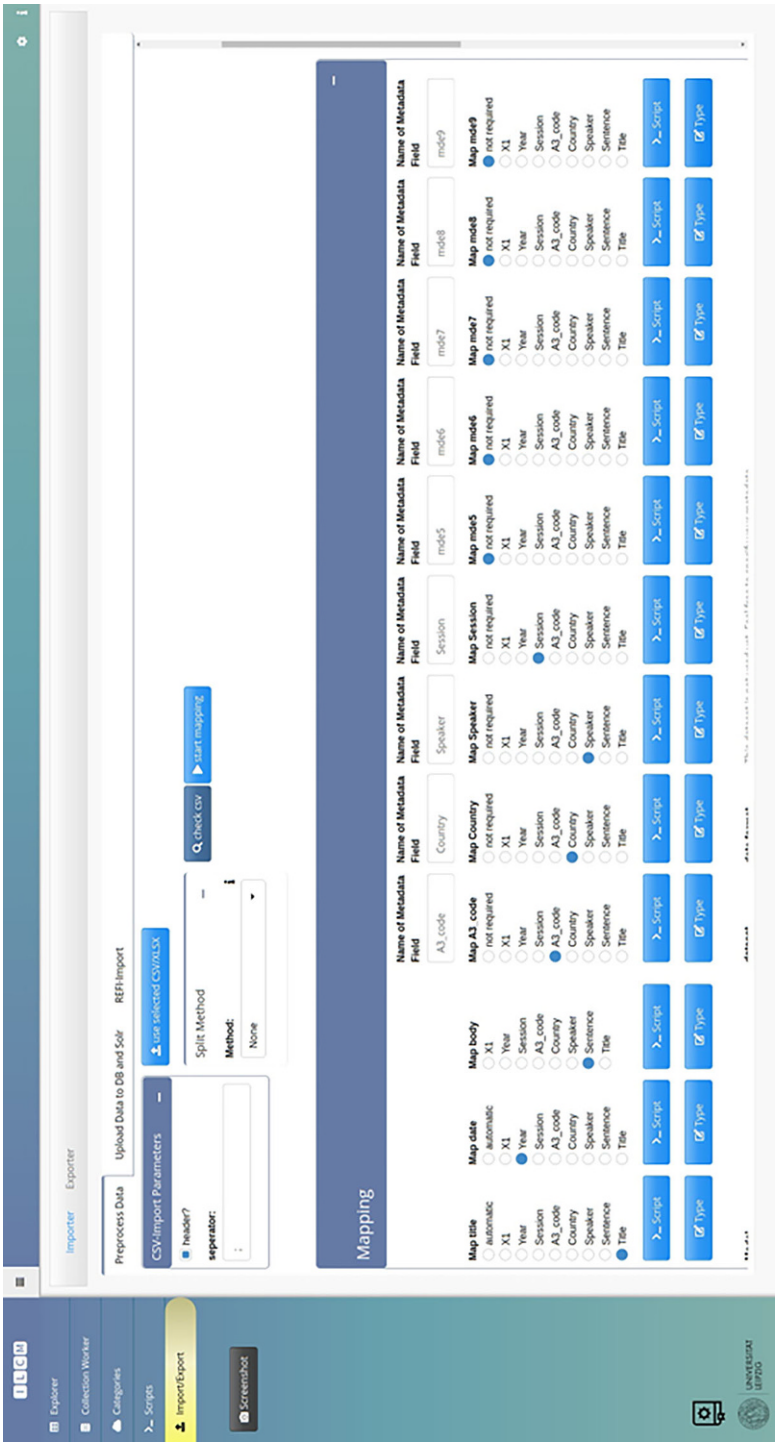


Fig. 11 Import interface of the iLCM, showing the mapping of the UNGA general debate CSV to the iLCM standard

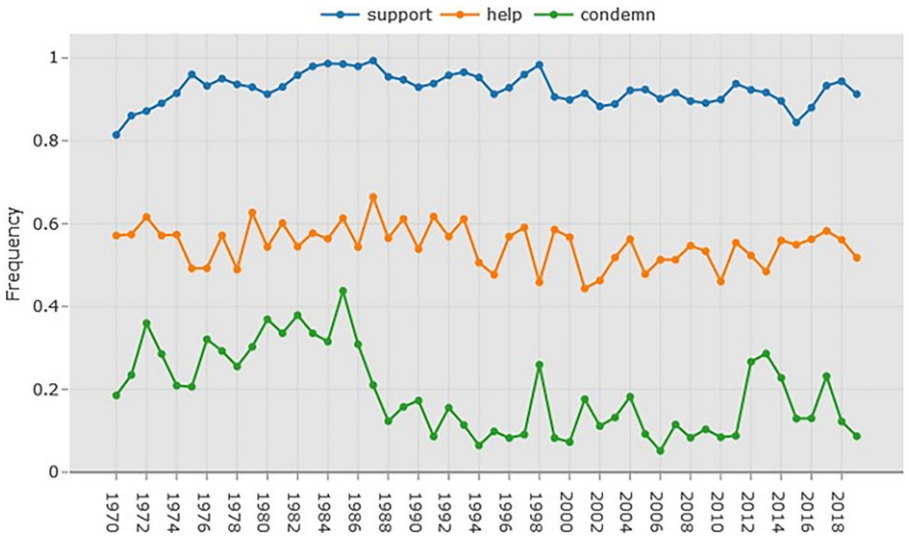


Fig. 12 Diachronic frequency plot of the verbs "support", "help", and "condemn"

3.2.2 Diachronic word frequencies

Based on the above qualitative findings, we can confidently state that, as would be expected of the UNGA, our corpus encompasses a notable quantity of expressive and declarative speech acts. This sets the stage for a subsequent quantitative analysis focusing on selected performative verbs and their probable associations with specific speech act types. For this purpose, we utilize the tool's frequency analysis function to investigate the diachronic frequencies of selected word forms. Figure 12 provides a snapshot of three performative verbs that indicate how declarative speech acts have evolved in the corpus over time.

3.2.3 Co-occurrence analysis

Furthermore, the use context of the performative verbs can be examined in more detail with the help of co-occurrence analysis to discover possible further indicators or peculiarities in the environment of the key words. For example, Fig. 13 shows words like "struggle", "full", "efforts", "community", and "assure" in semantic proximity to the performative verb "support". The graph also shows that there is a particularly strong relation between the words "international" and "community", which are likely an established bigram.

3.2.4 Supervised classification

The ultimate goal of this case study is to automatically identify and classify sentences according to different speech act types, so we can empirically investigate their distribution with respect to different speaker groups and time (who uses which

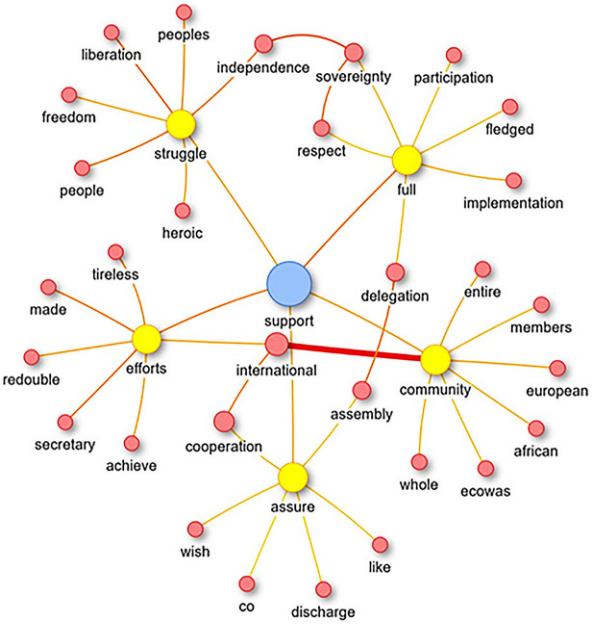


Fig. 13 Plot of the co-occurrence network for the term "support"

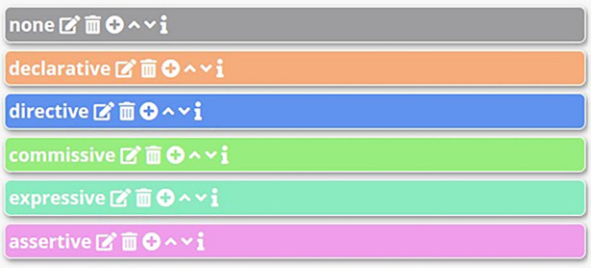


Fig. 14 Codebook created and used in the ILCM for the classification of speech acts

speak acts, and how does this change over time?). To achieve this, we first create a basic codebook with categories for each of the five different speech act types (see Fig. 14).

For the individual classes, references are then searched for in the corpus and annotated accordingly (see Fig. 15). Given that the process of annotating these speech acts can be time-consuming, we leverage the Active Learning feature of the iLCM to streamline the process. As noted in Sect. 2.4.5, AL involves the ability to train an initial classifier using manually annotated data, which is subsequently employed to present the user with new candidate examples for each class. This approach minimizes the manual annotation process, reducing it to evaluating the suggested samples. Furthermore, this iterative process can be repeated to achieve a robust classifier that can automatically annotate speech acts in all texts. These annotated speech acts can then be evaluated alongside other metadata for comprehensive analysis.

Figure 16 indicates that between 2002 and 2004 there is an increase in *declaratives* and a simultaneous decrease in *expressives*. From here, further qualitative

Fig. 15 Interface of the iLCM for displaying and annotating texts

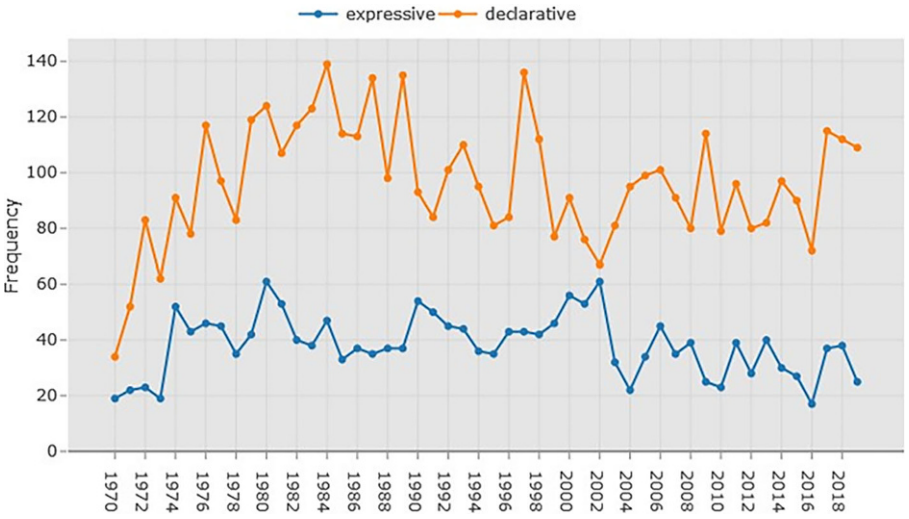


Fig. 16 Diachronic development of speech act types, declaratives and expressives, aggregated by year

and quantitative analysis can be performed, such as working with topic models to identify prevalent themes or subjects, analyzing named entitiesto extract and study specific entities mentioned in the text since they are already extracted by the linguistic preprocessing, or conducting a detailed analysis of individual documents.

3.2.5 Export of results

The results of the analysis can be extracted from the tool both in the form of example figures and as actual data in the form of CSV or R data objects. This data can then be further processed and analyzed in other environments (such as the RStudio IDE), which allows for a very flexible handling of diverse research questions and research hypotheses.

4 Conclusion, plans, and long-term support

In this article, we have discussed the current landscape of text analysis tools in the social sciences, highlighting a common trade-off between ease of use and static functionality, versus technical complexity with greater adaptability and flexibility. Existing tools tend to fall into one of these categories, and researchers are often faced with choosing between user-friendly but limited tools or more technically demanding but versatile ones. The iLCM aims to solve this problem by providing an easy-to-use GUI with a number of state-of-the-art text mining functions while simultaneously providing technically more advanced users with the opportunity to flexibly enhance the tool with further functions. Another key characteristic of the iLCM is its ability to seamlessly integrate qualitative approaches (such as annotation

and close reading) and quantitative approaches (various text mining methods) in a mixed-methods approach.

We would like to emphasize that one of the major strengths of the iLCM lies in its numerous functions in the realm of unsupervised and supervised learning, which Grimmer and Stewart (2013, pp. 267–297) highlight as crucial for computer-assisted text analysis, but which typically require advanced technical expertise to implement. We hope that the iLCM can contribute to the field of AI literacy and make such methods more accessible to a wider user base, enabling them to experiment with them in a user-friendly environment and to delve deeper into these techniques and expand upon them if desired, at a later stage.

To ensure the ongoing advancement of machine learning functionality, we are currently developing plans to integrate transformer-based large language models into the iLCM. Transformer-based deep learning methods have proven to be highly effective in processing and generating natural language text (see, for example, Lauriola et al. 2022; Sun et al. 2022). These methods employ mechanisms that enable them to capture long-range dependencies and contextual information within sequences of words, making them particularly advantageous for various text processing tasks, including language modeling, text classification, and even topic modeling. We are confident that incorporating these methods into tools like the iLCM will enable the realization of high-quality content-based research tasks in the future.

We also plan to enhance the functionality of the iLCM, allowing for seamless integration with existing research data infrastructures. This includes facilitating the import and export of data as well as ensuring sustainable documentation of results for reproducibility and intersubjective traceability, as these aspects have become crucial components of any research tool. Specifically, we have in mind the *Harvard Dataverse* (https://dataverse.harvard.edu/), *Zenodo* (https://zenodo.org/), *GitHub* (https://github.com/), or *Binder* (https://mybinder.org/), whose capabilities for storing data-based research for long-term archiving are expected to play an important role in realizing transparent research processes in the iLCM.

Funding This work was funded by Leipzig University and the Deutsche Forschungsgemeinschaft (Project number 324867496). The authors have no competing interests to declare that are relevant to the content of this article.

- ***Name of the resource:*** *interactive Leipzig Corpus Miner (iLCM)*
- ***Authors:*** *Andreas Niekler, Christian Kahmann, Manuel Burghardt, Gerhard Heyer*
- ***Link:*** *https://ilcm.informatik.uni-leipzig.de/*
- ***Description:*** *The iLCM project pursues the development of an integrated research environment for the analysis of structured and unstructured data in a "Software as a Service" architecture (SaaS). The research environment addresses requirements for the quantitative evaluation of large amounts of qualitative data using text mining methods and requirements for the reproducibility of data-driven research designs in the social sciences.*

  Das Projekt iLCM verfolgt die Entwicklung einer integrierten Forschungsumgebung für die Analyse von strukturierten und unstrukturierten Daten in einer 'Software as a Service'-Architektur (SaaS). Die Forschungsumgebung adressiert Anforderungen an die quantitative Auswertung großer Mengen qualitativer Daten mit Methoden des Text Mining und Anforderungen an die Reproduzierbarkeit datengetriebener Forschungsdesigns in den Sozialwissenschaften.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Anthony, L. (2005). Antconc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *IPCC 2005: Proceedings; International professional communication conference* (pp. 729–737). https://doi.org/10.1109/IPCC.2005.1494244.

Austin, J.L. (1962). *How to do things with words* (2nd edn.). Cambridge: Harvard University Press.

Baturo, A., Dasandi, N., & Mikhaylov, S.J. (2017). Understanding state preferences with text as data: introducing the UN General Debate Corpus. *Research and Politics*. https://doi.org/10.1177/2053168017712821.

Benko, V. (2013). Data deduplication in Slovak corpora. In K. Gajdošová & A. Žáková (Eds.), *Slovko 2013: Natural language processing, corpus linguistics, e-learning* (pp. 27–39). Lüdenscheid: RAM-Verlag.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774.

Biemann, C., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text: Eine Einführung in das Text Mining*. Wiesbaden: Springer Vieweg. https://doi.org/10.1007/978-3-658-35969-0.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Pytho*. Sebastpol: O'Reilly.

Blei, D.M., & Lafferty, J.D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). https://doi.org/10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What we can do and cannot do with topic modeling: a systematic review. *Communication Methods and Measures*, *17*(2), 111–130. https://doi.org/10.1080/19312458.2023.2167965.

Döveling, K., & Konijn, E. A. (Eds.). (2021). *Routledge international handbook of emotions and media*. New York: Routledge.

Dumouchel, S., Blotière, E., Barbot, L., Breitfuss, G., Chen, Y., Di Donato, F., Forbes, P., Petitfils, C., & Pohle, S. (2020). TRIPLE project: Building a discovery platform to enhance collaboration. *ITM Web of Conferences*, *33*, 3005. https://doi.org/10.1051/itmconf/20203303005.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*, 61–81. https://doi.org/10.1146/annurev-soc-121919-054621.

Früh, W. (2001). Kategorienexploration bei der Inhaltsanalyse. Basiswissengeleitete offene Kategorienbildung (BoK). In W. Wirth & E. Lauf (Eds.), *Inhaltsanalyse: Perspektiven, Probleme, Potentiale* (pp. 117–139). Köln: Halem.

Ganiger, S., & Rajashekharaiah, K. (2018). Comparative study on keyword extraction algorithms for single extractive document. In *2018 second international conference on intelligent computing and control cystems (ICICCS)* (pp. 1284–1287). https://doi.org/10.1109/ICCONS.2018.8663040.

Grimmer, J., & Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028.

Heyer, G., & Böhlke, V. (2021). CLARIN-D: an IT-based research infrastructure for the humanities and social sciences. In C. Koschtial, T. Köhler & C. Felden (Eds.), *E-science: open, social and virtual technology for research collaboration* (pp. 99–109). Cham: Springer. https://doi.org/10.1007/978-3-030-66262-27.

Heyer, G., Holz, F., & Teresniak, S. (2009). Change of topics over time—Tracking topics by their change of meaning. In *Proceedings of the international conference on knowledge discovery and information retrieval—KDIR (IC3K 2009)* (pp. 223–228). https://doi.org/10.5220/0002330602230228.

Hinrichs, E., Hinrichs, M., & Zastrow, T. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 25–29). https://aclanthology.org/P10-4005.

Honnibal, M., Boyd, A., & Montani, I. (2022). *spaCy: Industrial-strength natural language processing in Python* (Version v2.3.9). *Zenodo*. https://doi.org/10.5281/ZENODO.1212303.

Ignatow, G., & Mihalcea, R. (2016). *Text mining: a guidebook for the social sciences*. SAGE.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the Royal Society A: Mathematical, physical and engineering sciences*. https://doi.org/10.1098/rsta.2015.0202.

Kaggle (2022). Wikipedia movie plots. https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots. Accessed 31 Dec 2022.

Khoo, C. S., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. *Journal of Information Science*, *44*(4), 491–511. https://doi.org/10.1177/0165551517703514.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th edn.). Thousand Oaks: SAGE.

Kühne, R., Poggiolini, C., & Wirth, W. (2021). The differential effects of related and unrelated emotions on judgments about media messages. *Communications*, *46*(1), 127–149. https://doi.org/10.1515/commun-2020-2091.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing*, *470*, 443–456. https://doi.org/10.1016/j.neucom.2021.05.103.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742.

v d Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html.

Macanovic, A. (2022). Text mining for social science—The state and the future of computational text analysis in sociology. *Social Science Research*, *108*, 102784. https://doi.org/10.1016/j.ssresearch.2022.102784.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in

communication research: toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2), 93–118. https://doi.org/10.1080/19312458.2018.1430754.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). http://www.aclweb.org/anthology/P/P14/P14-5010.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv*. https://doi.org/10.48550/ARXIV.1301.3781.

Nabi, R. L. (2019). Media and emotion. In M. B. Oliver, A. A. Raney & J. Bryant (Eds.), *Media effects: Advances in theory and research* (4th edn., pp. 163–178). New York: Routledge.

Neuroth, H., Rapp, A., & Söring, S. (Eds.). (2015). *TextGrid: Von der Community – für die Community: Eine virtuelle Forschungsumgebung für die Geisteswissenschaften*. Göttingen: Universitätsverlag Göttingen. https://doi.org/10.3249/WEBDOC-3947.

Niekler, A., Wiedemann, G., & Heyer, G. (2014). Leipzig Corpus Miner—A text mining infrastructure for qualitative data analysis. In *Terminology and Knowledge Engineering 2014 (TKE 2014)*. https://hal.archives-ouvertes.fr/hal-01005878.

Niekler, A., Bleier, A., Kahmann, C., Posch, L., Wiedemann, G., Erdogan, K., Heyer, G., & Strohmaier, M. (2018). ILCM – A Virtual Research Infrastructure for Large-Scale Qualitative Data. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1209.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. University of Malta. (pp. 45–50). http://is.muni.cz/publication/884893/en.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, A. M., & Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*, 23. https://doi.org/10.1140/epjds/s13688-016-0085-1.

Richards, L. (2021). *Handling qualitative data: a practical guide* (4th edn.). Thousand Oaks: SAGE.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684.

Schröder, C., Niekler, A., & Potthast, M. (2022). Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2194–2203). https://doi.org/10.18653/v1/2022.findings-acl.172.

Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society*, *5*(1), 1–23.

Settles, B. (2012). *Active learning*. Springer. https://doi.org/10.1007/978-3-031-01560-1.

Simmler, S., Thorsten, V., & Pielström, S. (2019). *Topic modeling with interactive visualizations in a GUI Tool (Version V2)*. DataverseNL. https://doi.org/10.34894/ENV3TX.

Sun, T.-X., Liu, X.-Y., Qiu, X.-P., & Huang, X.-J. (2022). Paradigm shift in natural language processing. *Machine Intelligence Research*, *19*(3), 169–183. https://doi.org/10.1007/s11633-022-1331-6.

United Nations Framework Convention on Climate Change (2022). Nationally determined contributions (NDCs). https://unfccc.int/ndc-information/nationally-determined-contributions-ndcs. Accessed 31 Dec 2022.

Wiedemann, G. (2016). *Text mining for qualitative data analysis in the social sciences: a study on democratic discourse in Germany*. Wiesbaden: Springer.

Wiedemann, G., & Niekler, A. (2017). Hands-on: A five-day text mining course for humanists and social scientists in R. In *Proceedings of the Workshop on Teaching NLP for Digital Humanities (Teach4DH) co-located with GSCL 2017* (pp. 57–65). http://ceur-ws.org/Vol-1918/.

Wiedemann, G., Lemke, M., & Niekler, A. (2013). Postdemokratie und Neoliberalismus: Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949–2011 – ein Werkstattbericht. *ZPTh-Zeitschrift für Politische Theorie*, *4*(1), 99–115. https://www.budrich-journals.de/index.php/zpth/article/view/13868.

Dr. Andreas Niekler ist wissenschaftlicher Mitarbeiter in der Forschungsgruppe Computational Humanities am Institut für Informatik der Universität Leipzig.

Dr. Christian Kahmann ist wissenschaftlicher Mitarbeiter in der Forschungsgruppe Automatische Sprachverarbeitung an der Universität Leipzig

Dr. Manuel Burghardt ist Professor für Computational Humanities am Institut für Informatik der Universität Leipzig.

Dr. Gerhard Heyer ist Professor emeritus für Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig sowie externer Projektleiter für das Projekt NFDI Text+ an der Sächsischen Akademie der Wissenschaften.