Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230996

BUNDESTAG-MINE: Natural Language Processing for Extracting Key Information from Government Documents

Kevin BÖNISCH [a,1], Giuseppe ABRAMI [a],
Sabine WEHNERT [b,c] and Alexander MEHLER [a]
[a] *Goethe University Frankfurt, 60325 Frankfurt am Main, Germany*
[b] *Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany*
[c] *Leibniz Institute for Educational Media | GEI, 38118 Brunswick, Germany*
ORCiD ID: Kevin Bönisch https://orcid.org/0009-0005-6962-2646, Giuseppe Abrami https://orcid.org/0000-0002-7084-4909, Sabine Wehnert https://orcid.org/0000-0002-5290-0321, Alexander Mehler https://orcid.org/0000-0003-2567-7539

Abstract. As governments worldwide continue to release vast amounts of textual information, the need for efficient and insightful tools to extract, interpret and present this data has become increasingly critical. Towards solving this issue, we present the BUNDESTAG-MINE: an environment that periodically retrieves pertinent data from the German parliament, parses and analyzes it using pipelines for natural language processing, and then displays the results in a web application that is publicly accessible. BUNDESTAG-MINE helps to extract key information from parliamentary documents in a visually appealing matter for many use cases. For instance, the tool can be leveraged by journalists for news detection, lawyers for compliance checking, linguists for discourse analysis, and the broad public to inform themselves about the positions of political party members on a topic.

Keywords. political debate analysis, discourse visualization, sentiment analysis, named entity recognition, information retrieval

1. Introduction

Political and governmental actions have an impact on all areas of our daily lives by creating laws and regulations that both people and businesses must abide by. Following the most recent events and developments is therefore essential to not only react to, but also to foresee changes that may influence one's personal environment. This is particularly relevant for legal professionals, for example when checking compliance under changing regulatory conditions. To enable that, governments first document and then release all discussions, proposals and decisions in various formats for the general public to inter-

[1] Corresponding Author: Kevin Bönisch, k.boenisch@outlook.com

act with. Since the E-Government Act[2] entered into force in 2013, the German government is even obliged to publish this data. But analyzing the vast amount of data is not an easy task, as they are frequently not standardized, [1, p. 2], organized, or presented in an approachable or visually appealing manner. For solving these issues, *Natural Language Processing* (NLP) can be applied to standardize and sort the datasets, and to apply filters and algorithms to create meaningful statements and charts. Existing solutions have various limitations, e.g., the requirement for significant technological knowledge and the time commitment required to use them and to gain insights. To this end, we present BUNDESTAG-MINE, a platform for gathering German government data, having it analyzed using NLP techniques, and presenting the results in a user-friendly, publicly available web application[3].

Since rendering discourse in an understandable and non-cluttered manner is an active research field, it is crucial to look among related work. In doing so we find visualizations of keywords (top-n) [12], wordclouds [12], topic models [13], sentiment analysis [13], document similarity [13], discourse map from entities [2], and time series [12]. These works all focus on providing an overview of the main themes or the tone conveyed in the documents and are commonly used techniques in related tools which we will present in the following. An application for the Finnish Parliament is [9], who analyzes speeches with NLP and link them to a knowledge graph representing the activities of parliament members. Another tool for political discourse analysis is [4], who assign specific words to 30 different semantic classes. Furthermore, [5, 14] have analyzed parliamentary procedures for the German parliament since 1949. *SentiArt* [10, 11] analyze German political party programs, considering their semantic similarity and complexity, their main themes, the emotion potential and their readability.

2. Bundestags-Mine

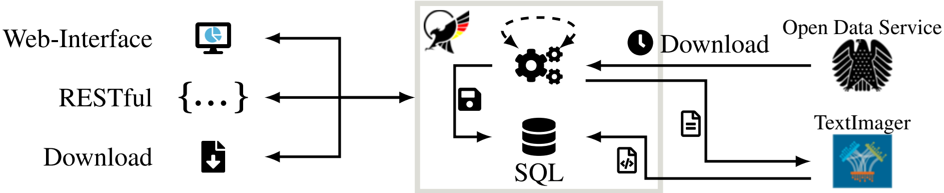
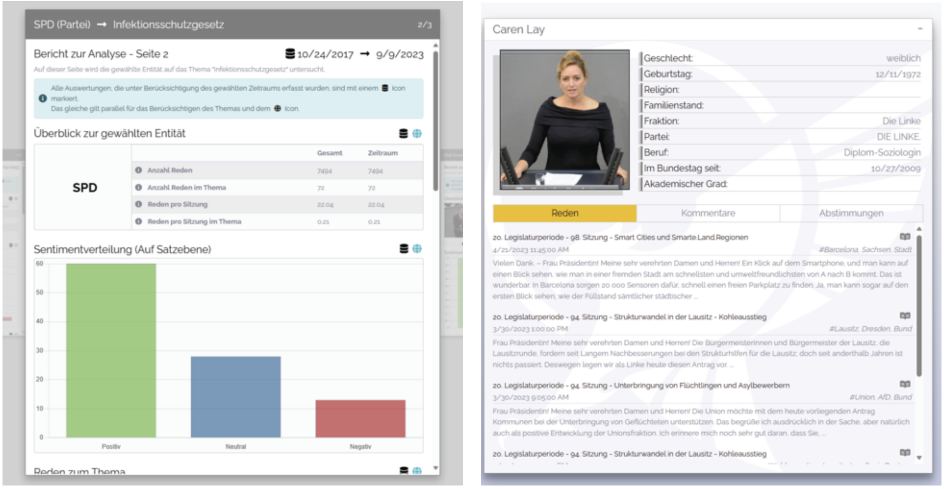


Figure 1. BUNDESTAG-MINE with its interfaces (left): Periodically, updates are downloaded from the *Open Data Service* of the German Bundestag and processed using *TextImager*; results are stored in an SQL database.

After understanding that no existing tool covers all features we require for a flexible analysis of the political discourse, we present the BUNDESTAG-MINE: an environment for evaluating German government documents by means of various Natural Language Processing techniques and visualizing them via a publicly accessable and intuitive web-application as well as providing the resulting data for download. Figure 1 provides a schematic overview of this work. The BUNDESTAG-MINE processes the following types

[2] Accessible here: https://tinyurl.com/ywznb9br
[3] https://bundestag-mine.de

of government data within its environment, which are then visualized within the web-application for a platform-independent and responsive interface: **minutes of plenary proceedings**, **agenda items** and **polls**. By analyzing this data, it is feasible to search for pertinent information about any governmental acts. For this purpose, BUNDESTAG-MINE uses the multi-server and multi-service system *TextImager* [7], a tool that annotates texts based on the *Unstructured Information Management applications* (UIMA) [6] standard. Using *spaCy* [8], the following annotations are accomplished: **Tokenization** divides the given text into smaller parts or tokens to prepare it for the application of further techniques. **Lemmatization** takes each token and determines its word stem. **Part-of-speech (POS) Tagging** determines the associated grammatical classification of each token like verb, adjective, adverb or noun. **Named-Entity-Recognition (NER)** involves detecting and categorizing important information in text known as named entities. These are then assigned into one of the four following groups: person, organisation, location and miscellaneous. By doing so we extract the relevant parts of the text, categorizing them and then using them for further analysis. Additionally, we use the **Sentiment Analysis**, based on [3], to determine the emotional tone of each sentence within each speech. Lastly, we **summarize** each speech with PEGASUS [16], a state-of-the-art transformer for abstract text summarization, and **translate** all German speeches to English via Opus-MT [15]. The data is made accessible for the user by a list of flexible, dynamic and parameterisable visualization features, implemented within the web-application to allow the actor to select, aggregate and interpret the data. These features, among others, include a **Search Engine** for fulltext search of speeches, comments, speakers, agenda items, and polls. The **Text Analysis** lists all parliamentary protocols together with the speakers, agenda items, polls, comments, and speeches for each protocol. The **Topic Analysis** (Figure 2a) which enables analysis of a specific speaker, party, or fraction in relation to a



SPD (Partei) → Infektionsschutzgesetz 2/3

Bericht zur Analyse - Seite 2 10/24/2017 → 9/9/2023

Auf dieser Seite wird die gewählte Entität auf das Thema "Infektionsschutzgesetz" untersucht.

Alle Auswertungen, die unter Berücksichtigung des gewählten Zeitraums erfasst wurden, sind mit einem Icon markiert.
Das gleiche gilt parallel für das Berücksichtigen des Themas und dem Icon.

Überblick zur gewählten Entität

		Gesamt	Zeitraum
SPD	Anzahl Reden	7494	7494
	Anzahl Reden im Thema	72	72
	Reden pro Sitzung	22.04	22.04
	Reden pro Sitzung im Thema	0.21	0.21

Sentimentverteilung (Auf Satzebene)

Caren Lay

Geschlecht:	weiblich
Geburtstag:	12/11/1972
Religion:	
Familienstand:	
Fraktion:	Die Linke
Partei:	DIE LINKE.
Beruf:	Diplom-Soziologin
Im Bundestag seit:	10/27/2009
Akademischer Grad:	

20. Legislaturperiode - 98. Sitzung - Smart Cities und Smarte.Land.Regionen
4/21/2023 11:45:00 AM #Barcelona, Sachsen, Stadt
Vielen Dank. – Frau Präsidentin! Meine sehr verehrten Damen und Herren! Ein Klick auf dem Smartphone, und man kann auf einen Blick sehen, wie man in einer fremden Stadt am schnellsten und umweltfreundlichsten von A nach B kommt. Das ist wunderbar. In Barcelona sorgen 20 000 Sensoren dafür, schnell einen freien Parkplatz zu finden. Ja, man kann sogar auf den ersten Blick sehen, wie der Füllstand sämtlicher städtischer ...

20. Legislaturperiode - 94. Sitzung - Strukturwandel in der Lausitz - Kohleausstieg
3/30/2023 1:00:00 PM #Lausitz, Dresden, Bund
Frau Präsidentin! Meine sehr verehrten Damen und Herren! Die Bürgermeisterinnen und Bürgermeister der Lausitz, die Lausitzrunde, fordern seit Langem Nachbesserungen bei den Strukturhilfen für die Lausitz; doch seit anderthalb Jahren ist nichts passiert. Deswegen legen wir als Linke heute diesen Antrag vor. ...

20. Legislaturperiode - 94. Sitzung - Unterbringung von Flüchtlingen und Asylbewerbern
3/30/2023 9:05:00 AM #Union, AfD, Bund
Frau Präsidentin! Meine sehr verehrten Damen und Herren! Die Union möchte mit dem heute vorliegenden Antrag Kommunen bei der Unterbringung von Geflüchteten unterstützen. Das begrüße ich ausdrücklich in der Sache, aber natürlich auch als positive Entwicklung der Unionsfraktion. Ich erinnere mich noch sehr gut daran, dass Sie. ...

20. Legislaturperiode - 94. Sitzung - Strukturwandel in der Lausitz - Kohleausstieg

(a) **Topic Analysis** about the party *SPD* connected to "Infection Protection Act"; sentiment analysis (🟩 positive, 🟦 neutral, 🟥 negative).

(b) The **Deputy Inspector** for Caren Lay, listing personal information alongside all of her speeches, comments and poll votes.

Figure 2. Two excerpts from the BUNDESTAG-MINE to visualize result details.

particular topic. And the **Deputy Inspector** (Figure 2b) to enable a closer examination of any deputy or speaker, providing information about their background, previous speeches, remarks, and poll results. Every feature in the BUNDESTAG-MINE is documented via video tutorials for better understanding and usage. Additionally, we offer a *Research Center* explaining the technological aspects of the environment and a *Download Center* which provides easy access to any pertinent data within BUNDESTAG-MINE.

References

[1] Giuseppe Abrami et al. "German Parliamentary Corpus (GerParCor)". In: *Proceedings of LREC*. Marseille, France: ELRA, 2022, pp. 1900–1906.

[2] Nicholas Botzer et al. "Entity graphs for exploring online discourse". In: *Knowl. Inf. Syst.* 65.9 (2023), pp. 3591–3609. DOI: 10.1007/s10115-023-01877-8.

[3] Mark Cieliebak et al. "A Twitter Corpus and Benchmark Resources for German Sentiment Analysis". In: *Proc. of SocialNLP*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 45–51. DOI: 10.18653/v1/W17-1106.

[4] Daniela Gîfu, et al. "Multi-dimensional analysis of political language". In: *Future Information Technology, Application, and Service: FutureTech 2012 Volume 1*. Springer. 2012, pp. 213–221.

[5] Limebit GmbH. *Open Discourse*. 2021. URL: https://opendiscourse.de/.

[6] T. Götz, et al. "Design and implementation of the UIMA Common Analysis System". In: *IBM Systems Journal* 43.3 (2004), pp. 476–489.

[7] Wahed Hemati et al. "TextImager: a Distributed UIMA-based System for NLP". In: *Proc. of COLING: System Demos*. Osaka, Japan, Dec. 2016, pp. 59–63.

[8] Matthew Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2015.

[9] Eero Hyvönen et al. "Plenary Speeches of the Parliament of Finland as Linked Open Data Data Services". In: *Proceedings of (ESWC 2023*. Ed. by Sanju Tiwari et al. Vol. 3447. CEUR Workshop Proceedings. CEUR-WS.org, 2023, pp. 1–20.

[10] Arthur M. Jacobs. "Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics". In: *Frontiers Robotics AI* 6 (2019), p. 53. DOI: 10.3389/frobt.2019.00053.

[11] Arthur M. Jacobs and Annette Kinder. "Electoral Programs of German Parties 2021: A Computational Analysis Of Their Comprehensibility and Likeability Based On SentiArt". In: *CoRR* abs/2109.12500 (2021). arXiv: 2109.12500.

[12] Paritosh D Katre. "NLP based text analytics and visualization of political speeches". In: *International Journal of Recent Technology and Engineering* 8.3 (2019), pp. 8574–8579.

[13] Salomon Orellana et al. "Using Natural Language Processing to Analyze Political Party Manifestos from New Zealand". In: *Inf.* 14.3 (2023), p. 152.

[14] Florian Richter et al. *Open Discourse: Towards the first fully Comprehensive and Annotated Corpus of the Parliamentary Protocols of the German Bundestag*. 2023.

[15] Jörg Tiedemann et al. "OPUS-MT – Building open translation services for the World". In: *Proc. of EAMT*. Lisboa, Portugal: EAMT, 2020, pp. 479–480.

[16] Jingqing Zhang et al. *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. 2019. DOI: 10.48550/ARXIV.1912.08777.