Topic modeling of parliamentary debates: comparing the state-of-the-art with embedding techniques for search engine applications

Submitted on: **19-02-2023**

Berend Driessen
berend.driessen@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Maarten Marx
m.j.marx@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Recent advancements of topic modeling techniques and an increase in the availability of data on European parliamentary proceedings opens up new research opportunities. One of these consists of generating insights through a parliamentary search engine based on topic modeling. This paper therefore examines the performance of topic modeling technique BERTopic on Dutch parliamentary proceedings. This model leverages embedding models and class-based TF-IDF to create interpretable topics. Performance evaluation consists of comparison with a baseline based on Non-Negative Matrix Factorization (NMF) and the use of topic coherence measures that correspond well with the human judgement of topic quality; UMass, UCI and NPMI. Finally, the paper demonstrates the potential for the implementation of this topic modeling technique for the enhancement of parliamentary search engines.

Keywords: Topic modeling, BERTopic, NMF, ParlaMint 2.1, parliamentary proceedings, search engine

1 INTRODUCTION

In today's age of information abundance, understanding the themes associated with a collection of text documents is a crucial challenge. Humans need powerful automatic tools that can effectively understand the underlying semantics of the data [14]. Topic modeling techniques are a suitable approach for this task. They consist of a group of algorithms that can discover latent topics within a text corpus by deriving document characteristics through the co-occurrences of words. Topic modeling has been used to identify the most prevalent themes in different settings, for example books, newspapers and social media posts [3]. These methods have also been employed in the political domain to analyze political attention by tracking and explaining policy agendas [8]. Ylä-Anttila, Eranti & Kukkonen show that topic modeling can be successfully used to track and compare the framing of a political issue, such as climate change, in the public debate between different countries [23]. However, one interesting application for topic modeling is the arena of political debates. This domain is especially suited for unveiling the political agenda, as politicians generally have limited time to express their views. Therefore, they have to decide what subjects to cover, which gives insights about their political priorities [8]. Recently, there has been an increase in the availability of data on political debates through the ParlaMint 2.1 project. This project consists of a multilingual collection of 17 comparable corpora containing parliamentary debates, with each corpus being about 20 million words in size [7]. The corpus stretches from 2015 to mid-2020 and contains rich metadata. This opens up potential for the task of topic modeling and topic analysis across different features.

As the number of applications for topic modeling increases, it becomes clear that not all topic modeling algorithms are well suited for all types of texts [3]. Factors to take into consideration are the length of the analyzed documents, text sparsity and volume. Furthermore, the performance of these algorithms can also be assessed based on their explanatory power in a social context [5]. This creates some inherent subjectivity to the evaluation of topic modeling techniques, even though the performance of models can be more objectively grasped through measuring the coherence of the presented topics [17]. Despite several weaknesses, Latent Dirichlet Allocation (LDA) enjoys popularity in the field and is often preferred for modeling topics. The drawbacks of LDA include requiring the number of topics to be known for optimal results and extra pre-processing tasks, such as custom stop-word lists, stemming and lemmatization [19]. Additionally, Non-Negative Matrix Factorization (NMF) is shown to be better suited to identify niche topics in specialized vocabularies, such as political debates [8]. However, just like LDA, NMF often relies on a bag-of-words representation of documents. This method ignores the ordering and semantics of words [1]. Consequently, recently emerged technologies such as Top2Vec [1] and BERTopic [10] are gaining popularity for their ability to capture this semantic information. Both models are proven to be capable of generating novel and interesting findings, and allow for a more in-depth understanding of certain topic based on search queries [5]. However, BERTopic is classified as having more potency for extracting useful information.

One domain that is underexposed in this specific context of topic modeling are the proceedings of Dutch parliamentary debates. Although this area has not been neglected scientifically, the focus of previous research lies on supervised sentimental analysis [9] or LDA-based topic modeling approaches [21]. However, the domain has not been exposed to the recent advances in the research field, such as the potential of using new techniques such as BERTopic. Furthermore, even though topic modeling can be used in the context of search engines [12], the concept of a search engine for parliamentary proceedings based on topic modeling has not yet been explored extensively.

Based on these observations, this papers proceeds to examine the following research question:

To what extent are embedded topic modeling techniques effective for generating insights when applied to parliamentary proceedings?

This research question is supported by several subquestions:

- How can topic modeling techniques be used to analyze parliamentary proceedings?
- How can the performance of topic modeling techniques be evaluated when applied to parliamentary proceedings?
- How easily can topic modeling of parliamentary proceedings be integrated in a parliamentary search engine, and how useful is this extra functionality?

2 RELATED WORK

Topic modeling is fast developing domain where new techniques are recently emerging. One of these techniques is BERTopic, which mpas word and sentence representation in vector space based on their textual embedding [10]. This enables capturing textual semantics, which leads to similar texts being close in vector space. Topics can be represented by clustering these embeddings with centroid-based techniques [18]. One technique to extract the topic representations is by finding the words that are in close proximity to a cluster's centroid. However, as a cluster will not always lie within a sphere around a cluster centroid, choosing words that are closest in proximity can lead to misleading topic representations [10]. Therefore, BERTopic utilizes a class-based version of TF-IDF to extract the topic representations from each topic. This should generally lead to better interpretable topics. However, a baseline needs to be established for evaluation. As suggested by Greene & Cross, NMF is effective in extracting topics from text with specialized vocabularies such as political debates [8]. This study will therefore use NMF as a baseline to evaluate the performance of the BERTopic model.

The unsupervised nature of topic modeling makes objective performance assessment a challenging endeavour. Perplexity has been a common method, but it is often negatively correlated with human judgements of topic quality [2]. However, there are still several ways in which the generated topics can be evaluated based on topic coherence that match well with human judgement of topic quality [20]. These topic coherence measures capture the semantic interpretability of discovered topics based on their corresponding descriptor terms [15]. For this study, we will therefore make use of the state-of-the-art topic coherence measures for evaluation, namely UMass, UCI and NPMI.

One of the applications of topic modeling is the enhancement of search engines [12]. According to Marti Hearst, the quality of a search engine is usually evaluated based on three main aspects of usability as defined by ISO 9241-11 [11]. These consists of the following criteria:

- Effectiveness: the accuracy and completeness with which users achieve specified goals.
- Efficiency: the resources expended in relation to the accuracy and completeness with which users achieve goals.
- Satisfaction: the freedom from discomfort, and positive attitudes towards the use of the product.

In short, it measures whether the users get an optimal experience when trying to retrieve relevant information based on their search queries. There is therefore a clear overlap with topic modeling; retrieving relevant and interpretable information from a certain dataset. Therefore, implementing topic modeling in a search engine is a logical extension of this technique and opens up possibilities for interactive human evaluations.

3 METHODOLOGY

The corpora used for analysis are part of the ParlaMint 2.1 project, and can be accessed through the corresponding CLARIN repository [6]. From the ParlaMint-NL files, the text (.txt) and meta documents (-meta.tsv) of all sessions from the House of Representatives (HoR) (Tweede Kamer) and Senate (Eerste Kamer) from 2015 to 2019 will be extracted. In each text document, all paragraphs of speech are on a separate line preceded by a code that acts as an ID, which is structured as follows: *ParlaMint-NL_YYYY-MM-DD-tweedekamer-X.uS*. The ID contains information about the date, whether the session belongs to the HoR or Senate, the session number and a reference number for the speaker. This ID corresponds to the ID in the document containing the meta information for each session. The metadata consists of information about the role of the speakers, their gender, their name, the party they belong to and whether they are part of the coalition or the opposition. As BERTopic uses sentence-transformers per default, it works best when documents consist of sentences or paragraphs. For working with BERTopic, we therefore want to create a dataframe containing individual speeches indexed by their ID. These ID's can later be used to trace back the sessions to the associated meta information. Therefore, we want to create two dataframes indexed by ID; one containing all the speeches called 'documents' and another containing all the meta-information called 'meta'. Each row in the 'documents' dataframe contains a separate speech, while each row in the 'meta' dataframe contains the information related to this session. Next, data exploration can be done to get a better understanding of the data.

3.1 Modelling process

As discussed earlier, BERTopic does not require further data preprocessing of the textual data. However, NMF does require further preprocessing. As mentioned earlier, BERTopic requires documents to consist of sentences or paragraphs, but this leads to many short documents because of the many non-informative interruptions present in the sessions. As NMF does not work well on short texts, all these short sentences need to be removed. After, further preprocessing is necessary such as the removal of stop words, tokenization, lowercasing and stemming. For the modelling itself, SciKit Learn will be used for NMF, so we can do feature creation based on TF-IDF, as this has been proven useful in the context of political texts [8]. The amount of topics will be selected based on maximizing UMass score. For BERTopic, the standard paraphrase-multilingual-MiniLM-L12-v2 embedding model is used as we are working the Dutch language.

3.2 Evaluation

Performance valuation of the topic modelling will be done by comparing the coherence of the generated topics by BERTopic and NMF. Three different measures will be used for this comparison, namely

UMass, UCI and NPMI. These different measures give a good indication whether the selected topics would score well on human-based interpretations of topic quality.

3.3 Search engine

The implementation of topic modeling in a search engine will be done through ElasticSearch. The aim is to implement the following functionalities:

- Search based on the search query's cosine similarity to the topics of speeches. Return highest scoring speeches in descending order of similarity.
- Filtering searches based on the metadata of individual speeches, such as the speaker, political parties, gender, coalition/opposition and year.

4 RISK ASSESSMENT

To assess the feasibility of the project, a thorough risk assessment is necessary. Therefore, potential obstacles will addressed in this section. First, a great risk for data science projects in general is related to obtaining the right data; that is, data that is sufficient in quantity and quality for the task at hand. For this study, the text corpus provided by ParlaMint 2.1 does seem to fulfill both these criteria. However, further data exploration is needed for confirmation. Additionally, as only unsupervised techniques are employed, there is no need for the notoriously time-consuming task of annotating data. Second, using unsupervised models does create a new problem for researchers. As these models are not trained on labeled data, their results can not be compared to some golden standard or ground truth. This introduces fuzziness and subjectivity in performance evaluation. To create solid, informative results, the evaluative criteria need to be well-defined and well-argued for. Therefore, this research will use several metrics for performance as discussed in the related works section. Third, there might risks associated with the process, such as modelling and the implementation in a search engine. This covers risks related to technical and time-based feasibility. To address this risk in an early stage, the focus will lie on laying down a clear path to execution in an early stage to assess feasibility.

Although the aim is to mitigate these potential risks, a Plan B is deemed necessary to increase the chances for finishing the thesis project in time. This Plan B consists of two alternative setups for the research design that are a more explicit continuation of the existing research on parliamentary proceedings. If the risks of the original design unfold into insurmountable obstacles, the Plan B forms a concrete backup plan that enables the project to say afloat. The setups are based on insights from three different papers. First, a paper by Kerkvliet, Kamps & Marx that discuss the performance of the named entity recognition (NER) technique SpaCy when applied to Dutch parliamentary proceedings [13]. For future work, they suggest comparing the performance of SpaCy with using of transfer learning techniques with unsupervised learned embeddings like BERT or ELMO for NER on Dutch proceedings. This suggestion is grounded in previous work which shows that these models can outperform state-of-the-art NER approaches [16][4]. The first alternative setup therefore consists of comparing the performance of BERT for the NER taks on the labelled Dutch parliament corpus from this paper. A potential research question for this research design could be formulated as follows:

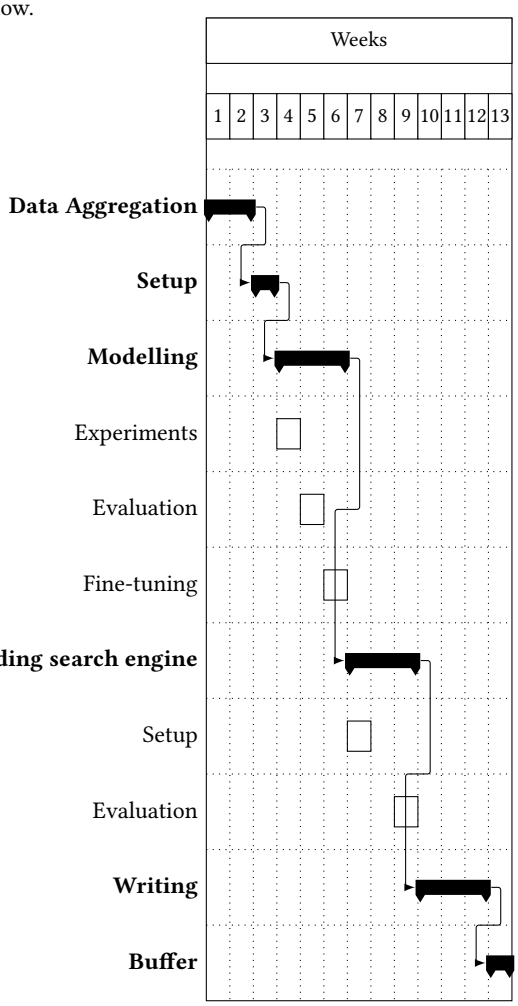
To what extent are unsupervised learned embeddings using transferable learning effective and accurate in entity recognition techniques when applied to parliamentary proceedings?

Second, a paper by van Heusden, Kamps & Marx that proposes a coreference resolution for Dutch parliament proceedings, and compares its performance to SoNaR-1 and RiddleCoref, two other existing Dutch coreference resolution corpora [22]. In the section discussing future work, they propose investigating the addition of structured metadata from structured files into the neural models. A potential research question for this research design could be formulated as follows:

To what extent does the addition of structured metadata improve the performance of coreferencing techniques when applied to parliamentary proceedings?

5 PROJECT PLAN

The proposed timeline of the project can be found in the Gantt chart below.



REFERENCES

[1] Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470* (2020).

[2] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems* 22 (2009).

[3] Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *Comput. Surveys* 54, 10s (2022), 1–35.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology* 7 (2022).

[6] Tomaž Erjavec, Vladislava Grigorova, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Andrej Pančur, Michał Rudolf, and Kiril Simov. 2020. Multilingual comparable corpora of parliamentary debates ParlaMint 1.0. (2020).

[7] Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Dargis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. 2021. Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. http://hdl.handle.net/11356/1432 Slovenian language resource repository CLARIN.SI.

[8] Derek Greene and James P Cross. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis* 25, 1 (2017), 77–94.

[9] Steven Grijzenhout, Maarten Marx, and Valentin Jijkoun. [n. d.]. Sentiment analysis in parliamentary proceedings. ([n. d.]).

[10] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[11] Marti Hearst. 2009. *Search user interfaces*. Cambridge university press.

[12] Abdulrahman Helan and Zainab Namh Sultani. 2023. Topic modeling methods for text data analysis: A review. In *AIP Conference Proceedings*, Vol. 2457. AIP Publishing LLC, 040006.

[13] Lennart Kerkvliet, Jaap Kamps, and Maarten Marx. 2020. Who mentions whom? recognizing political actors in proceedings. In *Proceedings of the second parlaclarin workshop*. 35–39.

[14] Pooja Kherwa and Poonam Bansal. 2019. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems* 7, 24 (2019).

[15] Derek O'callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications* 42, 13 (2015), 5645–5657.

[16] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474* (2019).

[17] Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. 2014. Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397* (2014).

[18] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914* (2020).

[19] Raquel Silveira, C Fernandes, João A Monteiro Neto, Vasco Furtado, and José Ernesto Pimentel Filho. 2021. Topic modelling of legal documents via legal-bert. *Proceedings http://ceur-ws org ISSN* 1613 (2021), 0073.

[20] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 952–961.

[21] Janneke M van der Zwaan, Maarten Marx, and Jaap Kamps. 2016. Validating Cross-Perspective Topic Modeling for Extracting Political Parties' Positions from Parliamentary Proceedings.. In *ECAI*. 28–36.

[22] Ruben van Heusden, Jaap Kamps, and Maarten Marx. 2023. Neural Coreference Resolution for Dutch Parliamentary Documents with the DutchParliament Dataset. *Data* 8, 2 (2023), 34.

[23] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication* 18, 1 (2022), 91–112.