Semantic Similarity of Parliamentary Speech using BERT Language Models & fastText Word Embeddings

Katja Meden
Department of Knowledge Technologies E8,
Jožef Stefan Institute
katja.meden@ijs.si

ABSTRACT

The main objective of this paper is to present the work done on comparing the two methods for measuring semantic similarity of parliamentary speech between coalition and opposition regarding the adoption of the first COVID-19 epidemic response package. We first measured sentence similarity using four BERT-based language models (Language agnostic BERT Sentence Encoder - LaBSE model, Sentence-LaBSE, Sentence-BERT, multilingual BERT - mBERT) and compared the results amongst them. Using the word embedding method, fastText, we then measured the semantic similarity of full-text parliamentary speech and presented the results using descriptive analysis. Lastly, we compared the usage of both methods and highlighted some of the advantages and disadvantages of each method for measuring the semantic similarity of parliamentary speech.

KEYWORDS

parliamentary speech, semantic similarity, sentence similarity, BERT language models, fastText

1 INTRODUCTION

"National parliamentary data is a verified communication channel between the elected political representatives and society members in any democracy. It needs to be made accessible and comprehensive - especially in times of a global crisis." [13] In parliamentary discourse, politicians expound their beliefs and ideas through argumentation and to persuade the audience, they highlight some aspect of an issue. If we are to understand the role of parliamentary discourse practices, we need to explore the recurring linguistic patterns and rhetorical strategies used by MPs that help to reveal their ideological commitments, hidden agendas, and argumentation tactics [11]. One of the ways to study the aforementioned linguistic patterns can be done by researching similarities of parliamentary speeches using different methods for measuring semantic similarity of text.

The aim of this paper is to present the work done on comparing the two methods for measuring semantic similarity of parliamentary speech between coalition and opposition regarding the adoption of the first COVID-19 epidemic response package. We measured sentence similarity with four BERT-based language models (Language agnostic BERT Sentence Encoder - LaBSE model [7], Sentence-LaBSE [8], Sentence-BERT [14], multilingual BERT – mBERT [1]) and compared the scores of most similar and least similar sentences.

To facilitate the intended scope of our initial research, i.e., researching similarity of full-text parliamentary speech, we used fastText [5] and presented results using descriptive analysis to gain additional insight into the characteristics of coalition and opposition parliamentary speech. Lastly, we highlighted some of the advantages and disadvantages of each method for measuring semantic similarity of parliamentary speech.

The paper is structured as follows: Section 2 contains an overview of the related work on word embeddings and language models. Section 3 presents the methodology and we describe the experiment setting in Section 4. The experiment results are found in Section 5. Finally, we conclude the paper and provide ideas for future work in Section 6.

2 RELATED WORK

Two blocks of texts are considered similar if they contain the same words or characters. Techniques like Bag of Words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF) can be used to represent text as real value vectors to aid calculation of Semantic Textual Similarity (STS) [3]. STS is defined as the measure of semantic equivalence between two blocks of text and usually give a ranking or percentage of similarity between texts, rather than a binary decision as similar or not similar [3]. Word embeddings are one of the methods developed to aid in measuring semantic similarity. They provide vector representations of words where vectors retain the underlying linguistic relationship between the similarities of the words. Word embeddings consist of two types: static and contextualized word embeddings. With static word embeddings, words will always have the same representation, regardless of the context where it occurs, while with contextualized word embedding, representation depends on the context of where that word occurs – meaning, that the same word in different contexts can have different representations.

FastText is an open-source, free, lightweight library that allows users to learn text representations and text classifiers [5]. It is a representative of the static word embedding technique, where a vector representation is associated to each character n-gram; words being represented as the sum of these representations [2]. The fundamental problem of word embeddings is that they generate the same embedding for the same word in different contexts, failing to capture polysemy [4].

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia

Language models are contextualized word representations that aim at capturing word semantics in different contexts to address the issue of polysemy and the context of words [4]. BERT, or Bidirectional Encoder Representations from Transformer, is a language model, designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers [6]. BERT word representations are therefore contextual

3 METHODOLOGY

3.1 Dataset

Dataset contains 230 documents (speeches) from the Extraordinary Session 33 from the corpora of the Slovenian parliamentary debates (ParlaMint-SI) [9] from 2014 to mid-2020, linguistically annotated and represented in the CoNNL-U format (which include POS, lemmatized and NER tags**)**. We chose an extraordinary session in a time of crisis for two reasons: firstly, regular sessions deal with multiple problems (such as MP questions), which makes a comparison between speeches difficult. Similarly, we chose only one specific theme (the adoption of the first epidemic response package), which helped in the initial analysis and comparison of documents.

3.2 Data analysis and pre-processing

For the initial data analysis, we used the Orange data-mining tool [12] that helped us with the data understanding and initial dataset pre-processing.

For full speech measuring with fastText we removed speeches by Chairperson to avoid adding noise to the dataset in the form of procedural speech that would make measuring semantic similarities almost impossible. We also removed Slovene stopwords and manually added a list of four additional stopwords: *hvala*, *danes*, l*epa* and *beseda*, which excluded the very common phrase *Hvala za besedo* (eng. Thank you for the word) and its variations. Some of the documents were missing the *party_status* labels (values: *coalition* and *opposition*). The missing values (17 documents) were thus removed from the dataset. The pre-process gave us a total of 97 documents, presented in Table 1. Looking at the distributions of the speeches in the session, almost 1/3 of the speeches belongs to the opposition. Both coalition and opposition consists of four political parties: LMŠ, Levica, SAB and SD are part of the opposition, all of mostly left and centre-left political orientation. Similarly, the coalition consists of DeSUS, NSi, SDS and SMC political parties[1], all mostly right-winged and centre-right parties.

Table 1: Preprocessed dataset

Sample	Number of documents	Total
Coalition	30 (30.93%)	97
Opposition	67 (69.07%)	

[1] Technically, the opposition consists of 5 political parties, but SNS (Slovenska Nacionalna Stranka) does not have any speeches in the dataset.

We used the same settings for the second part of the experiment (comparing sentence similarity with the four BERT-based models) with one difference. Since all BERT-based models support *max_lenght* input in the size 512 tokens, we decided to filter out sentences that refer explicitly to the response package (keyword for selection being *zakon*). To facilitate the visualisations and balance out our dataset, we randomly chose 20 sentences for each group (coalition/opposition).

3.3 Experiment settings

As mentioned, BERT-based models have restrictions on the maximum length of input documents. For most, this is 512 tokens, and in the case of Sentence-BERT, this restriction is even more severe (128 word tokens). Most speeches in the dataset are longer than the maximum length – this limitation did not allow us to conclude semantic similarity measurement on full parliamentary speech. The first part of the experiment therefore focuses on sentence similarity. From previously described BERT-based models, three of the models were fine-tuned for sentence similarity tasks: Sentence-LaBSE [7], LaBSE [8], mBERT [1] and Sentence-BERT [14]. For easier comparison, we used mean pooling and cosine distance to measure the similarity.

To achieve the intended scope of our initial research (researching the semantic similarity of parliamentary speech), we used the fastText-based Orange widget *Document embedding* (using mean as the aggregation method) to embed our documents and calculate cosine similarity to achieve comparison between coalition and opposition parliamentary speech. With these two experiments, we can compare measuring semantic similarity with language models to the word embedding method (fastText). This comparison would be better with Longformer language model (which can take up to around 1000+ word tokens as *max_input*) as we could compare methods for measuring semantic similarity of full-text documents (speeches), but as of time of writing this paper, Longformer [10] does not yet support Slovene language.

4 RESULTS

4.1 Results of the sentence similarity measure with BERT-based models

As stated previously, we used four different BERT-based models to measure semantic similarity of 40 sentences (20 sentences for each group - coalition and opposition) and visualized the results using heat maps (example in Figure 1). Initially, we first selected well-known BERT-based models that were optimized for Slovene (trilingual model CroSloEngual BERT and monolingual model SloBERTa), that did not produce reliable results - as shown in Table 2, CroSloEngual [15] and SloBERTa [16] produce extremely high similarity scores, since, as we later discovered, were not fine-tuned for sentence similarity task.

Table 2: Similarity scores of language models for most similar and least similar sentences

Model	Most similar	Least similar
Sentence-LaBSE	0.6184	0.1235
LaBSE	0.7610	0.3649
mBERT	0.8930	0.5377
Sentence-BERT	0.6677	-0.0792
CroSloEngual	0.9931	0.9480
SloBERTa	0.9867	0.8899

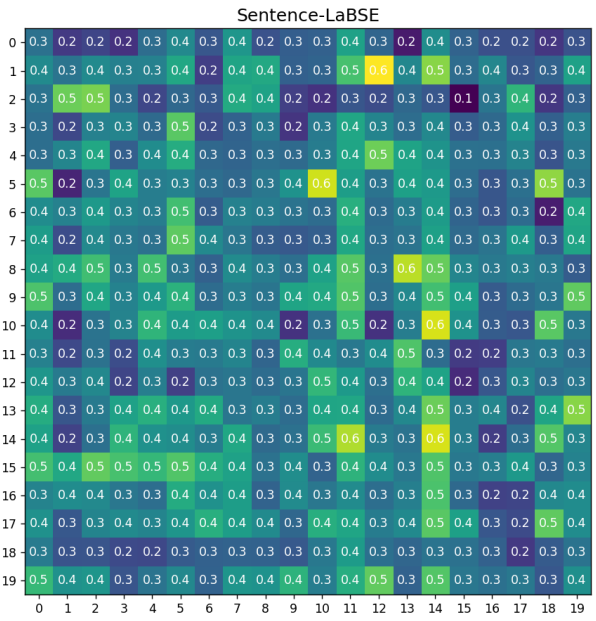

Figure 1: Example of heat map using Sentence-LaBSE model

When comparing the models, it does not surprise that Sentence-LaBSE and Sentence-BERT show very similar results (see Table 2), as they come from the same family of models and thus have similar model architecture (and are both fine-tuned for this specific task). What is interesting is the fact that Sentence-BERT is the only model that produced a negative score for the least similar sentence (similarity score of -0.0792), while mBERT model showed the highest similarity scores (outside of CroSloEngual and SloBERTa). Some of the highest scored sentences showed that speakers from different party statuses tend to use similar language patterns, for example:

Coalition*: "Ob hitrem sprejemanju zakona je potrebno zagotoviti, da ne bodo spregledane posamezne ranljive skupine posameznikov."*
(Eng. "With the rapid adoption of the law, it is necessary to ensure that individual vulnerable groups of individuals are not overlooked.")

Opposition*: "Še enkrat, ostaja še cela vrsta ranljivih skupin v zakonu, ki je nenaslovljena."*
(Eng. "Once again, there is a whole range of vulnerable groups in the law that remain unaddressed.")

4.2 Results of the document similarity with fastText

For the second part of our experiment, we used fastText for word embedding and measured cosine distance to get semantic similarity score of our documents. Figure 2 shows visualized results comparing speeches between coalition and opposition speakers:

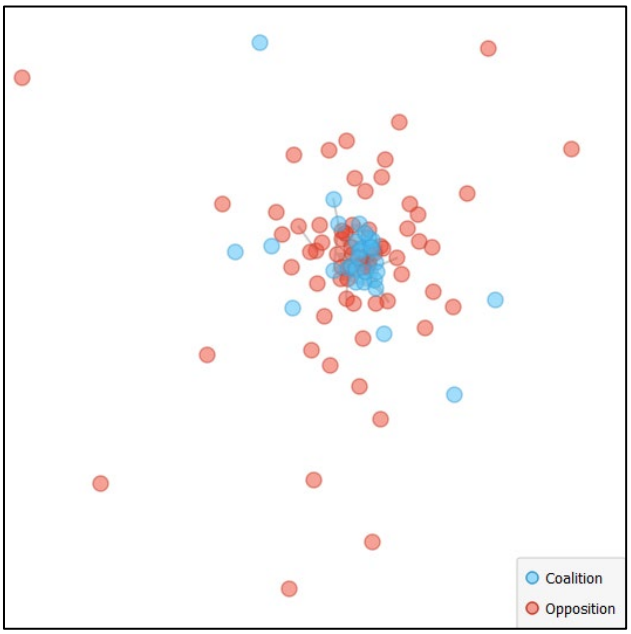

Figure 2: Document similarity with fastText, visualized using MDS

Documents (or speeches) are connected closely together – this could be attributed mostly to the fact that they are addressing the same issue – the adoption of the first epidemic release package. The most similar speeches were made by the members of the political party SDS (coalition) and SD (opposition), followed closely by SMC and Levica. All speeches are long and focus on the topic of the session – the proposed law (most speeches include keywords such as "*zakon*" (law), "*zakonski paket*" (law packet), "*amandma*" (amendment), "*ukrepi*" (measures).

Outlier detection analysis showed 8 speeches (7 made by the opposition, 1 by coalition), which are all very short and focus solely on parliamentary procedures. We also observed some trends in the usage of the words, concatenated from the word "*korona*": "*koronakriza*", "*koronazakon*", "*antikoronazakon*", "*koronaobveznica*", "*koronapomoči*", "*protikoronapaket*" etc. (used mostly by the opposition).

In Figure 3, we compared speech between the members of the opposition. The visualization showed a cluster of similar speeches. Members of Levica seemed to be most vocal during the session (by having more than 50% of all opposition speeches), while also having several similar speeches, with the central sub-topic being proposed amendments to the law and financial consequences of it. The least similar speech was made by Violeta Tomić, member of Levica, in regard to the date the epidemic was declared.

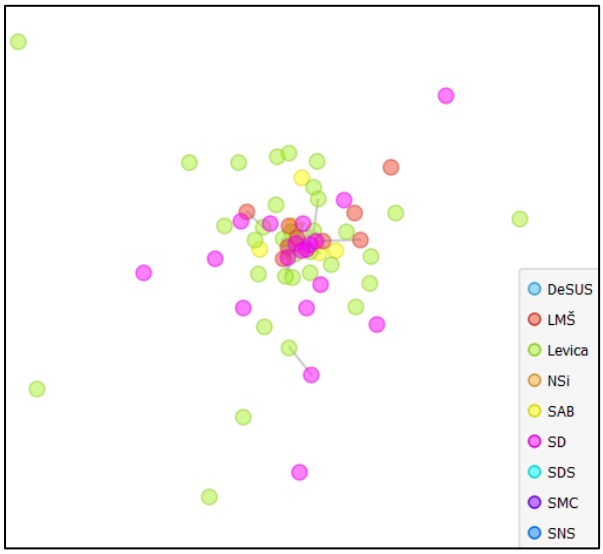

Figure 3: Document similarity with fastText (opposition)

In Figure 4, we compared speech between the members of coalition: speeches are less connected; with most similar divided among SDS members, closely connected to the SMC, NSI and DeSUS members. The common sub-topic to all of the speeches made is the financial crisis as a direct result of the epidemic. Two of the most far-away speeches belong to the member of DeSUS (Franc Jurša). Both speeches are among the shortest ones in the dataset, with a focus on the topic of pensions and registration of a parliamentary group, and thus are not explicitly connected to the central topic of the discourse.

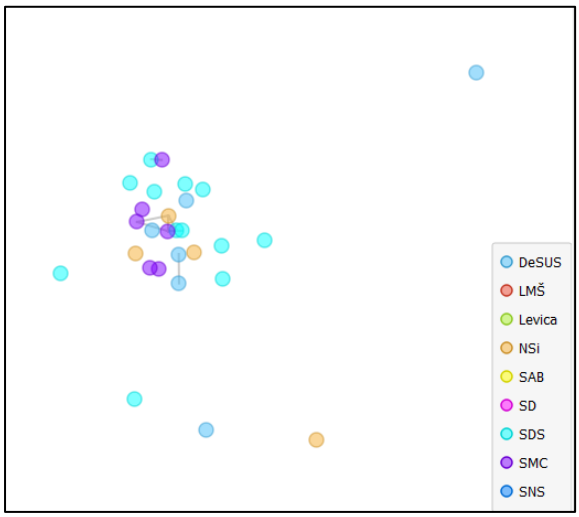

Figure 4: Document similarity with fastText (coalition)

5 CONCLUSIONS

In this paper, we were comparing language models and word embeddings as methods for measuring semantic similarity of parliamentary speech. In the initial stages, it turned out that there is not a lot of models that support Slovene as input language. Those that were made explicitly with Slovene in mind (such as SloBERTa and CroSloEngual BERT) were not fine-tuned for semantic similarity/sentence similarity tasks and thus do not produce accurate results. Limitation on maximum length of input text that most BERT-based models have is probably one of the biggest disadvantages of the language models for semantic similarity measures (this is being alleviated with new emerging language models, such as Longformer, that allow over 1000+ tokens as maximum input length). For sentence similarity task language models from Sentence-BERT family show the most accuracy and are easier to use as standard BERT models (such as mBERT).

Even though BERT contextualizes word embeddings (and therefore might produce better results because of it), fastText solved the problem of text-input length and combined with Orange data mining tool allowed us to explore similarities between speeches as we originally intended to do. From the document similarity analysis, we saw that most speeches were relatively connected (similar) to one another. Speeches amongst the members of the opposition were more similar in comparison to the speeches made amongst coalition members. There were a few outlier speeches in both opposition and coalition – they were all shorter speeches and less related to the original topic of the discourse. For future work, some limitations of this research should first be addressed (e.g., comparing language models to word embedding techniques on a full-text basis) and repeat the experiments with fine-tuned SloBERTa and CroSloEngual model on full ParlaMint-SI corpora.

REFERENCES

[1] BERT multilingual base model (cased): https://huggingface.co/bert-base-multilingual-cased

[2] Bojanowski, Piotr, Grave, Edouard, Joulin, Armand and Mikolov, Tomas. (2017). Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, 5, 135-146. DOI: https://doi.org/10.1162/tacl_a_00051

[3] Chandrasekaran, Dhivya, and Vijay Mago. 2021. Evolution of Semantic Similarity—A Survey. In *ACM Computing Surveys,* 1-37.

[4] David S. Batista. 2018. Language Models and Contextualised Word Embeddings. https://www.davidsbatista.net/blog/2018/12/06/Word_Embeddings/

[5] FastText - Library for efficient text classification and representation learning. https://fasttext.cc/

[6] 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*.

[7] Language-agnostic BERT Sentence Encoder (LaBSE) (Sentence-Transformers): https://huggingface.co/sentence-transformers/LaBSE

[8] Language-agnostic BERT Sentence Encoder (LaBSE): https://huggingface.co/setu4993/LaBSE

[9] Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. 2021. http://hdl.handle.net/11356/1431

[10] Longformer: https://huggingface.co/docs/transformers/model_doc/longformer

[11] Naderi, Nona, and Graeme Hirst. 2015. Argumentation mining in parliamentary discourse. In Principles and practice of multi-agent systems, 16-25. https://cmna.csc.liv.ac.uk/CMNA15/paper%209.pdf

[12] Orange: Data Mining Tool for visual programming. https://orangedatamining.com/

[13] ParlaMint: Towards Comparable Parliamentary Corpora. 2020. https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

[14] Sentence-BERT (sentence-transformers/distiluse-base-multilingual-cased-v2): https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

[15] Ulčar, Matej and Robnik-Šikonja, Marko, 2020, *CroSloEngual BERT 1.1*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1330.

[16] Ulčar, Matej and Robnik-Šikonja, Marko, 2021, *Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1397