The ParlaSent-BCS Dataset of Sentiment-annotated Parliamentary Debates from Bosnia and Herzegovina, Croatia, and Serbia

Michal Mochtak,[*] Peter Rupnik,[†] Nikola Ljubešić[†‡]

[*]Institute of Political Science
University of Luxembourg
2 avenue de l'Université, L-4366 Esch-sur-Alzette
michal.mochtak@uni.lu

[†] Department of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
peter.rupnik@ijs.si
nikola.ljubesic@ijs.si

[‡]Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, SI-1000 Ljubljana

Abstract

Expression of sentiment in parliamentary debates is deemed to be significantly different from that on social media or in product reviews. This paper adds to an emerging body of research on parliamentary debates with a dataset of sentences annotated for detection of sentiment polarity in political discourse using sentence-level data. We sample the sentences for annotation from the proceedings of three Southeast European parliaments: Croatia, Bosnia and Herzegovina, and Serbia. A six-level annotation schema is applied to the data with the aim of training a classification model for the detection of sentiment in parliamentary proceedings. Krippendorff's alpha measuring the inter-annotator agreement ranges from 0.6 for the six-level annotation schema to 0.75 for the three-level schema and 0.83 for the two-level schema. Our initial experiments on the dataset show that transformer models perform significantly better than those using a simpler architecture. Furthermore, regardless of the similarity of the three languages, we observe differences in performance across different languages. Performing parliament-specific training and evaluation shows that the main reason for the differing performance between parliaments seems to be the different complexity of the automatic classification task, which is not observable in annotator performance. Language distance does not seem to play any role neither in annotator nor in automatic classification performance. We release the dataset and the best-performing models under permissive licences.

1. Introduction

Emotions and sentiment in political discourse are deemed as crucial and influential as substantive policies promoted by the elected representatives (Young and Soroka, 2012). Since the golden era of research on propaganda (Lasswell, 1927; Shils and Janowitz, 1948), a number of scholars have demonstrated the growing role of emotions on affective polarization in politics with negative consequences for the stability of democratic institutions and the social cohesion (Garrett et al., 2014; Iyengar et al., 2019; Mason, 2015). With the booming popularity of online media, sentiment analysis has become an indispensable tool for understating the positions of viewers, customers, but also voters (Soler et al., 2012). It has allowed all sorts of entrepreneurs to know their target audience like never before (Ceron et al., 2019). Experts on political communication argue that the way we receive information and how we process them play an important role in political decision-making, shaping our judgment with strategic consequences both on the level of legislators and the masses (Liu and Lei, 2018). Emotions and sentiment simply do play an important role in political arenas and politicians have been (ab)using them for decades.

Although there is a general agreement among political scientists that sentiment analysis represents a critical component for understanding political communication in general (Young and Soroka, 2012; Flores, 2017; Tumasjan et al., 2010), the empirical applications outside the English-speaking world are still rare (Rauh, 2018; Mohammad, 2021). This is especially the case for studies analyzing political discourse in low-resourced languages, where the lack of out-of-the-box tools creates a huge barrier for social scientists to do such research in the first place (Proksch et al., 2019; Mochtak et al., 2020; Rauh, 2018). The paper, therefore, aims to contribute to the stream of applied research on sentiment analysis in political discourse in low-resourced languages. The goal is to present a new annotated dataset compiled for machine-learning applications focused on the detection of sentiment polarity in the political discourse of three Southeast European (SEE) countries: Bosnia and Herzegovina, Croatia, and Serbia. We further use the dataset to train different classification models for the sentiment analysis applying different schemas and settings to demonstrate the benefits and limitations of the dataset and the trained models. We release the dataset and the best-performing models under permissive licenses to facilitate

further research and more empirically oriented projects. In general, the paper, the dataset, and the models contribute to an emerging community of research outputs on parliamentary debates with a focus on sentence-level sentiment annotation with future downstream applications in mind.

2. Dataset construction

2.1. Focus on sentences

The dataset we compile and then use for training different classification models focuses on a sentence-level data and utilizes sentence-centric approach for capturing sentiment polarity. The strategy goes against the tradition in mainstream research applications in social sciences which focus either on longer pieces of text (e.g. utterance of "speech segment" or whole documents (Bansal et al., 2008; Thomas et al., 2006)) or coherent messages of shorter nature (e.g. tweets (Tumasjan et al., 2010; Flores, 2017)). The approach, however, creates certain limitations when it comes to political debates in national parliaments where speeches range from very short comments counting only a handful of sentences to long monologues having thousands of words. Moreover, as longer text may contain a multitude of sentiments, any annotation attempt must generalize them, introducing a complex coder bias which is embedded in any subsequent analysis. The sentence-centric approach attempts to refocus the attention on individual sentences capturing attitudes, emotions, and sentiment positions and using them as lower-level indices of sentiment polarity in a more complex political narrative. Although sentences cannot capture complex meanings as paragraphs or whole documents do, they usually carry coherent ideas with relevant sentiment affinity. This approach stems from a tradition of content analysis in political science which focuses both on the political messages and their role in political discourse in general (Burst et al., 2022; Hutter et al., 2016; Koopmans and Statham, 2006).

Unlike most of the literature which approaches sentiment analysis in political discourse as a proxy for position-taking stances or as a scaling indicator (Abercrombie and Batista-Navarro, 2020b; Glavaš et al., 2017; Proksch et al., 2019), a general sentence-level classifier we aim for in this paper has a more holistic (and narrower) aim. Rather than focusing on a specific policy or issue area, the task is to assign a correct sentiment category to sentence-level data in political discourse with the highest possible accuracy. Only when a good performing model exists, a downstream task can be discussed. We believe it is a much more versatile approach which opens a wide range of possibilities for understanding the context of political concepts as well as their role in political discourse. Furthermore, sentences as lower semantic units can be aggregated to the level of paragraphs or whole documents which is often impossible the other way around (document $\rightarrow$ sentences). Although sentences as the basic level of analysis are less common in social sciences research when it comes to computational methods (Abercrombie and Batista-Navarro, 2020b), practical applications in other areas exist covering topics such as validation of sentiment dictionaries (Rauh, 2018), ethos mining (Duthie and Budzynska, 2018), opinion mining (Naderi and Hirst, 2016), or detection of sentiment carrying sentences (Onyimadu et al., 2013).

2.2. Background data

In order to compile a dataset of political sentiment for manual annotation and then use it for training the classification models for real world applications, we sampled sentences from three corpora of parliamentary proceedings in the region of former Yugoslavia – Bosnia and Herzegovina (Mochtak et al., 2022c),[1] Croatia (Mochtak et al., 2022a),[2] and Serbia (Mochtak et al., 2022b).[3] The Bosnian corpus contains speeches collected on the federal level from the official website of the Parliamentary Assembly of Bosnia and Herzegovina (Parlamentarna skupština BiH, 2020). Both chambers are included – House of Representatives (Predstavnički dom / Zastupnički dom) and House of Peoples (Dom naroda). The corpus covers the period from 1998 to 2018 (2nd – 7th term) and counts 127,713 speeches. The Croatian corpus of parliamentary debates covers debates in the Croatian parliament (Sabor) from 2003 to 2020 (5th – 9th term) and counts 481,508 speeches (Hrvatski sabor, 2020). Finally, the Serbian corpus contains 321,103 speeches from the National Assembly of Serbia (Skupština) over the period of 1997 to 2020 (4th – 11th term) (Otvoreni Parlament, 2020).

2.3. Data sampling

Each speech was processed using the CLASSLA-Stanza tool (Ljubešić and Dobrovoljc, 2019) with tokenizers available for Croatian and Serbian in order to extract individual sentences as the basic unit of our analysis. In the next step, we filtered out only sentences presented by actual speakers, excluding moderators of the parliamentary sessions. All sentences were then merged into one meta dataset. As we want to sample what can be understood as "average sentences", we further subset the sentence meta corpus to only sentences having the number of tokens within the first and third frequency quartile (i.e. being within the interquartile range) of the original corpus ($\sim$3.8M sentences). Having the set of "average sentences", we used the Croatian gold standard sentiment lexicon created by (Glavaš et al., 2012), translated it to Serbian with a rule-based Croatian-Serbian translator (Klubička et al., 2016), combined both lexicons, and extracted unique entries with a single sentiment affinity, and used them as seed words for sampling sentences for manual annotation. The final pool of seed words contains 381 positive and 239 negative words (neutral words are excluded). These seed words are used for stratified random sampling which gives us 867 sentences with negative seed word(s), 867 sentences with positive seed word(s), and 866 sentences with neither positive nor negative seed words (supposedly having neutral sentiment). We sample 2600 sentences in total for manual annotation. The only strata we use is the size of the original corpora (i.e. number of sentences per corpus). With this we sample 1,388 sentences from the Croatian parliament, 1059

[1] https://doi.org/10.5281/zenodo.6517697
[2] https://doi.org/10.5281/zenodo.6521372
[3] https://doi.org/10.5281/zenodo.6521648

sentences from the Serbian parliament, and 153 sentences from the Bosnian parliament.

2.4. Annotation schema

The annotation schema for labelling sentence-level data was adopted from Batanović et al. (Batanović et al., 2020) who propose a six-item scale for annotation of sentiment polarity in a short text. The schema was originally developed and applied to SentiComments.SR, a corpus of movie comments in Serbian and is particularly suitable for low-resourced languages. The annotation schema contains six sentiment labels (Batanović et al., 2020: 6):

- +1 (`Positive` in our dataset) for sentences that are entirely or predominantly positive
- –1 (`Negative` in our dataset) for sentences that are entirely or predominantly negative
- +M (`M_Positive` in our dataset) for sentences that convey an ambiguous sentiment or a mixture of sentiments, but lean more towards the positive sentiment in a strict binary classification
- –M (`M_Negative` in our dataset) for sentences that convey an ambiguous sentiment or a mixture of sentiments, but lean more towards the negative sentiment in a strict binary classification
- +NS (`P_Neutral` in our dataset) for sentences that only contain non-sentiment-related statements, but still lean more towards the positive sentiment in a strict binary classification
- –NS (`N_Neutral` in our dataset) for sentences that only contain non-sentiment-related statements, but still lean more towards the negative sentiment in a strict binary classification

The different naming convention we have applied in our dataset serves primarily practical purposes: obtaining the 3-way classification by taking under consideration only the second part of the string (if an underscore is present).

Additionally, we also follow the original schema which allowed marking text deemed as sarcastic with a code "sarcasm". The benefit of the whole annotation logic is that it was designed with versatility in mind allowing reducing the sentiment label set in subsequent processing if needed. That includes various reductions considering polarity categorization, subjective/objective categorization, change of the number of categories, or sarcasm detection. This is important for various empirical tests we perform in the following sections.

2.5. Data annotation

Data were annotated in two waves, with 1300 instances being annotated in each. Annotation was done via a custom online app. The first batch of 1300 sentences was annotated by two annotators, both being native speakers of Croatian, while the second batch was annotated only by one of them.

parliament	positive	neutral	negative
all	470	772	1358
HR	261	433	694
BS	27	42	84
SR	182	297	580

Table 1: Distribution of the three-class labels in the whole dataset, as well as across each of the three parliaments.

The inter-annotator agreement (IAA) measured using Krippendorff's alpha in the first round was 0.599 for full six-item annotation scheme, 0.745 for the three-item annotation schema (positive/negative/neutral), and 0.829 for the two-item annotation schema focused on the detection of only negative sentiment (negative/other). The particular focus on negative sentiment in the test setting is inspired by a stream of research in political communication which argues that negative emotions appear to be particularly prominent in the context of forming the human psyche and its role in politics (Young and Soroka, 2012). More specifically, political psychologists have found that negative political information has a more profound effect on attitudes than positive information as it is easier to recall and is more useful in heuristic cognitive processing for simpler tasks (Baumeister et al., 2001; Utych, 2018).

Before the second annotator moved to annotate the second batch of instances, hard disagreements, i.e. disagreements pointing at a different three-class sentiment, where +NS and -NS are considered neutral, were resolved together by both annotators through a reconciliation procedure.

The final distribution of the three-class labels in the whole dataset, as well as along specific parliaments, is given in Table 1. The presented distributions show that, regardless of a lexicon-based sampling, the negative class is still by far the most pervasive category, which might be even more the case in a randomly sampled dataset, something we leave for future work.

2.6. Dataset encoding

The final dataset, available through the CLARIN.SI repository, contains the following metadata:

- `sentence` that is annotated
- `country` of origin of the sentence
- annotation `round` (first, second)
- annotation of `annotator1` with one of the labels from the annotation schema presented in Section 2.4.
- annotation of `annotator2` following the same annotation schema
- annotation given during `reconciliation` of hard disagreements
- the three-way `label` (positive, negative, neutral) where +NS and -NS labels are mapped to the neutral class

- the `document_id` the sentence comes from
- the `sentence_id` of the sentence
- the `date` the speech was given
- the `name`, `party`, `gender`, `birth_year` of the speaker
- the `split` (train, dev, or test) the instance has been assigned to (described in more detail in Section 3.1.

The final dataset is organized in a JSONL format (each line in the file being a JSON entry) and is available under the CC-BY-SA 4.0 license.[4]

3. Experiments

3.1. Data splits

For performing current and future experiments, the dataset was split into the train, development and test subsets. The development subset consists of 150 instances, while the test subset consists of 300 instances, both using instances from the first annotation round, where two annotations per instance and hard disagreement reconciliations are available. The training data consists of the remainder of the data from the first annotation round and all instances from the second annotation round, summing to 2150 instances.

While splitting the data, stratification was performed on the variables of three-way sentiment, country, and party. With this we can be reasonably sure that no specific strong bias regarding sentiment, country or political party is present in any of the three subsets.

3.2. Experimental setup

In our experiments we investigate the following questions: (1) how well can different technologies learn our three-way classification task, (2) what is the difference in performance depending on which parliament the model is trained or tested on, and (3) is the annotation quality of the best performing technology high enough to be useful for data enrichment and analysis.

We investigate our first question by comparing the results on the following classifiers: fastText (Joulin et al., 2016) with pre-trained CLARIN.SI word embeddings (Ljubešić, 2018), the multilingual transformer model XLM-Roberta (Conneau et al., 2019),[5] the transformer model pre-trained on Croatian, Slovenian and English cseBERT (Ulčar and Robnik-Šikonja, 2020),[6], and the transformer model pre-trained on Croatian, Bosnian, Montenegrin and Serbian BERTić (Ljubešić and Lauc, 2021).[7] Our expectation is for the last model to perform best given that it was pre-trained on most data from the three languages. However, this assumption has to be checked given that for some tasks even models pre-trained on many languages obtain performance that is comparable to otherwise superior models pre-trained on one or few languages (Kuzman et al., 2022).

model	macro F1
classla/bcms-bertic	**0.7941 ± 0.0101****
EMBEDDIA/crosloengual-bert	0.7709 ± 0.0113
xlm-roberta-base	0.7184 ± 0.0139
fasttext + CLARIN.SI embeddings	0.6312 ± 0.0043

Table 2: Results of the comparison of various text classification technologies. We report macro-F1 mean and standard deviation over 6 runs with the model-specific optimal number of training epochs. The distributions of results of the two best performing models are compared with the Mann-Whitney U test (** $p < 0.01$).

While comparing the different classification techniques, each model was optimized for the epoch number hyperparameter on the development data, while all other hyperparameters were kept default. For training transformers, the simpletransformers library[8] was used.

The second question on parliament specificity we answer by training separate models on Croatian sentences only and Serbian sentences only, evaluating each model both on Croatian and on Serbian test sentences. We further evaluate the model trained on all training instances separately on instances coming from each of the three parliaments.

For our third question on the usefulness of the model for data analysis, we report confusion matrices, to inform potential downstream users of the model's per-category performance.

4. Results

4.1. Classifier comparison

We report the results of our text classification technology comparison in Table 2. The results show that transformer models are by far more capable than the fasttext technology relying on static embeddings only. Of the three transformer models, the multilingual XLM-RoBERTa model shows to have a large gap in performance to the two best-performing models. Comparing the cseBERT and the BERTić model, the latter manages to come on top with a moderate improvement of 1.5 points in macro-F1. The difference in the results of the two models is statistically significant regarding the Mann-Whitney U test (Mann and Whitney, 1947), with a p-value of 0.0053.

4.2. Parliament dependence

We next investigate the dependence of the results on from which parliament the training and the testing data came. Our initial assumption was that the results are dependent on whether the training and the testing data come from the same or a different parliament, with same-parliament results being higher. We also investigate how the model trained on all data performs on parliament-specific test data.

[4] http://hdl.handle.net/11356/1585

[5] https://huggingface.co/xlm-roberta-base

[6] https://huggingface.co/EMBEDDIA/crosloengual-bert

[7] https://huggingface.co/classla/bcms-bertic

[8] https://simpletransformers.ai

4.2.1. Impact of training data

We perform this analysis on all three transformer models from Section 4.1., hoping to obtain a deeper understanding of parliament dependence on our task. We train and test on data from the Croatian and the Serbian parliament only as the Bosnian parliament's data are not large enough to enable model training.

In Table 3 we report the results grouped by model and training and testing parliament. To our surprise, the strongest factor shows not to be whether the training and testing data come from the same parliament, but what testing data are used, regardless of the training data. This trend is to be observed regardless of the model used.

The results show that Serbian test data seem to be harder to classify, regardless of what training data are used, with a difference of ∼9 points in macro-F1 for the BERTić and the XLM-RoBERTa models. The difference is smaller for the cseBERT model, ∼7 points, but still shows the same trend as the two other models.

We have additionally explored the possibility of a complexity bias of Serbian test data in comparison to Serbian training data by performing different data splits, but the results obtained were very similar to those presented here. Serbian data seem to be harder to classify in general, which is observed when performing inference over Serbian data. Training over Serbian data still results in a model comparably strong to that based on Croatian training data. Important to note is that the Croatian data subset is 30% larger than the Serbian one.

To test whether the Serbian data complexity goes back to challenges during data annotation, or whether it is rather the models that struggle with inference over Serbian data, we calculated the Krippendorff IAA on data from each parliament separately. The agreement calculation over the ternary classification schema resulted in an IAA for Bosnian data of 0.69, Croatian data of 0.733, and Serbian data of 0.77. This insight proved that annotators themselves did not struggle with Serbian data as these had the highest IAA. We also tested whether there is excessive sarcasm in Serbian data, which might affect the model's performance. The dataset contains two sarcastic instances from the parliament of Bosnia and Herzegovina and 16 for both Croatia and Serbia, which means sarcasm can hardly explain the overall lower performance on Serbian test data. Lastly, we checked the type-token ratio (TTR) on samples of Croatian and Serbian sentences to estimate the lexical richness of each subset, a higher lexical richness of Serbian (via a higher type-token ratio) possibly explaining the lower results obtained on Serbian test data. By calculating the type-token ratio on 100 tokens selected from random sentences, and repeating the process 100 times in a bootstrapping manner, we obtained a result of 0.833 for Serbian and 0.839 for Croatian. This result shows for the Croatian part of the dataset to be just slightly more lexically rich (83.9 different tokens among 100 tokens on average) than Serbian (83.3 different tokens among 100 tokens), which does not explain the difference in performance of various classifiers on Serbian data.

The complexity of Serbian data that can be observed in the evaluation is due to some effect that we did not manage to identify at this point, but that will have to be taken under consideration in future work on this dataset.

XLM-RoBERTa		
train \test	HR	SR
HR	0.7296 ± 0.0251	0.6128 ± 0.0341
SR	0.7323 ± 0.0282	0.6487 ± 0.0203
cseBERT		
train \test	HR	SR
HR	0.7748 ± 0.0174	0.7146 ± 0.0175
SR	0.7762 ± 0.0114	0.6989 ± 0.0275
BERTić		
train \test	HR	SR
HR	0.8147 ± 0.0083	0.7249 ± 0.0105
SR	0.7953 ± 0.0207	0.7130 ± 0.0278

Table 3: Comparison of the three transformer models when trained and tested on data from the Croatian or Serbian parliament. Average macro-F1 and standard deviation over 6 runs is reported.

test	ternary	binary
all	0.7941 ± 0.0101	0.8999 ± 0.0120
HR	0.8260 ± 0.0186	0.9221 ± 0.0153
BS	0.7578 ± 0.0679	0.9071 ± 0.0525
SR	0.7385 ± 0.0170	0.8660 ± 0.0150

Table 4: Average macro-F1 and standard deviation of 6 runs of the BERTić model, trained on all training data, and evaluated on varying testing data.

4.2.2. Impact of testing data

In the next set of experiments, we compare the performance of BERTić classifiers trained over all training data, but evaluated on all and per-parliament testing data. Beyond this, we train models over the ternary schema that we have used until now (positive vs. neutral vs. negative), but also the binary schema (negative vs. rest), given our special interest in identifying negative sentences, as already discussed in Section 2.5.

We report results on test data from each of the three parliaments, including the Bosnian one, which, however, contains only 18 testing instances, so these results have to be taken with caution.

The results presented in Table 4 show again that the Serbian data seem to be the hardest to classify even when all training data are used. Bosnian results are somewhat close to the Serbian ones, but caution is required here due to the very small test set. This level of necessary caution regarding Bosnian test data is also visible from the five times higher standard deviation in comparison to the results of the two other parliaments. Croatian data seem to be easiest to classify, with an absolute difference of 9 points between the performance on Serbian and Croatian test data. Regarding the binary classification results, these are, as expected, higher than those of the ternary classification schema with an macro-F1 of 0.9 when all data are used. The relationship between specific parliaments is very similar to that observed using the ternary schema.

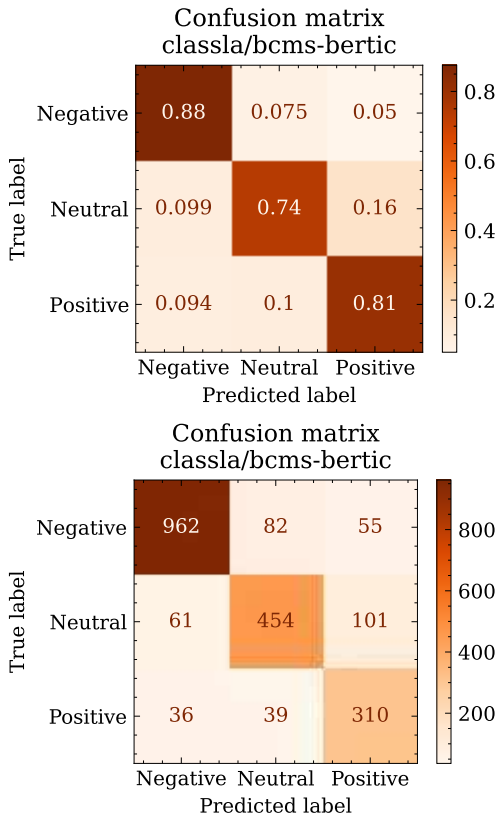

Figure 1: Row-normalised and raw-count confusion matrix of the BERTić results on the ternary schema.

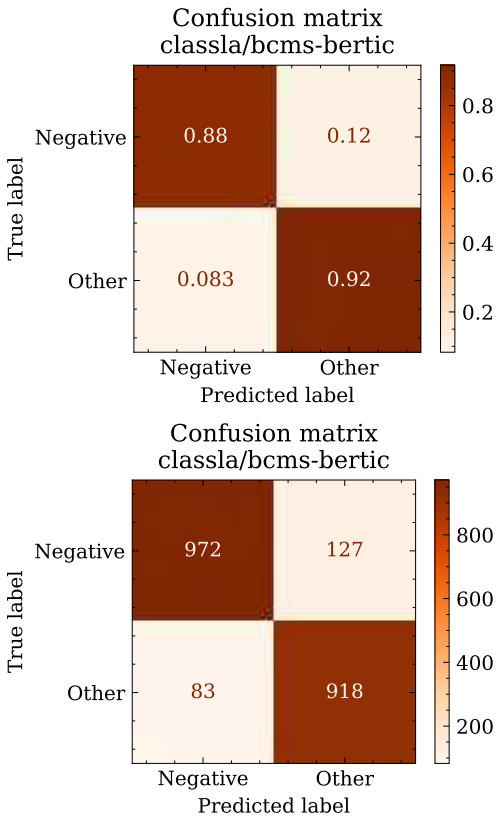

Figure 2: Row-normalised and raw-count confusion matrix of the BERTić results on the binary schema.

4.3. Per-category analysis

Our final set of experiments investigates the per-category performance both on the ternary and the binary classification schema. We present the confusion matrices on the ternary schema, one row-normalized, another with raw counts, in Figure 1. As anticipated, the classifier works best on the negative class, with 88% of negative instances properly classified as negative. Second by performance is the positive class with 81% of positive instances being labelled like that, while among the neutral instances 3 out of 4 instances are correctly classified. Most of the confusion between classes occurs, as expected, between the neutral and either of the two remaining classes.

The binary confusion matrices, presented in Figure 2 show for a rather balanced performance on both categories. On each of the categories recall is around 0.9, with a similar precision given the symmetry of the confusions.

When comparing the output of the ternary and the binary model, the ternary model output mapped to a binary schema performs slightly worse than the binary model, meaning that practitioners should apply the binary model if they are interested just in distinguishing between negative and other sentences.

Although any direct comparisons are hard to make, the few existing studies which performed text classification on sentence-level data, report much worse results. Rauh (2018) found that when three annotators and three sentiment dictionaries were compared on a ternary classification task (positive/negative/neutral), they agreed only in one-quarter of the 1,500 sentences. Using heuristic classifiers based on the use of statistical and syntactic clues, Onyimadu et al. (2013) found that on average, only 43% of the sentences were correctly annotated for their sentiment affinity. The results of our experiments are therefore certainly promising. Especially when it comes to the classification of negative sentences, the model has 1 in 10 sentence error rate which is almost on par with the quality of annotation performed by human coders.

5. Conclusion

The paper introduces a sentence-level dataset of parliamentary proceedings, manually annotated for sentiment via a six-level schema. The good inter-annotator agreement is reported, and the first results on the automation of the task are very promising, with a macro-F1 of ∼0.8 on the ternary schema and ∼0.9 on the binary schema. The difference in performance across the three parliaments is observed, but visible only during inference, Serbian data being harder to make predictions on, while for modelling, all parliaments seem to be similarly useful. One limitation of our work is the following: our testing data have been sampled as the whole dataset, with a bias towards mid-length sentences, and sentences containing sentiment words. Future work should consider preparing a sample of random sentences, or, even better, consecutive sentences, so that the potential issue of lack of a wider context during manual data annotation is successfully mitigated as well.

In general, the reported results have several promising implications for applied research in political science. First of all, it allows a more fine-grained analysis of political concepts and their context. A good example is a combination of the KWIC approach with sentiment analysis, with a focus on examining the tone of a message in political discourse. This is interesting for both qualitatively and quantitatively oriented scholars. Especially the possibility of extracting numeric assessment of the classification model (e.g. class probability) is particularly promising for all sorts of hypothesis-testing statistical models. Moreover, sentence-level analysis can be combined with the findings of various information and discourse theories for studying political discourse focused on rhetoric and narratives (e.g. beginning and end of a speech are more relevant than what comes in the middle). Apart from the concept-driven analysis, the classification model can be used for various research problems ranging from policy position-taking to ideology detection or general scaling tasks (Abercrombie and Batista-Navarro, 2020a; Glavaš et al., 2017; Proksch et al., 2019). Although each of these tasks requires proper testing, the performance of the trained models for such applications is undoubtedly promising.

As a part of our future work, we plan to test the usefulness of the predictions on a set of downstream tasks. The goal is to analyze the data from all three parliaments (Bosnia and Herzegovina, Croatia, and Serbia) in a series of tests focused on replication of the results from the existing research using mostly English data. Given the results we obtained, we aim to continue our research using the setup with the model trained on cross-country data. Furthermore, the three corpora we have used in this paper will be extended as a part of ParlaMint II project.

We make the ternary and binary BERTić models trained on all available training available via the HuggingFace repository[9][10] and make the dataset available through the CLARIN.SI repository (Mochtak et al., 2022d).

Acknowledgements

This work has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278341 (MaCoCu project). This communication reflects only the author's view. The Agency is not responsible for any use that may be made of the information it contains.

This work was also funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programme "Language resources and technologies for Slovene" (P6-0411).

6. References

Gavin Abercrombie and Riza Batista-Navarro. 2020a. ParlVote: A corpus for sentiment analysis of political debates. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Gavin Abercrombie and Riza Batista-Navarro. 2020b. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*, 3(1):245–270.

Mohit Bansal, Claire Cardie, and Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the Min-cut classification framework. In: *Coling 2008: Companion volume: Posters*, pages 15–18, Manchester, UK. Coling 2008 Organizing Committee.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. 2020. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLOS ONE*, 15(11):e0242050.

Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad is Stronger than Good. *Review of General Psychology*, 5(4):323–370.

Tobias Burst, Werner Krause, Pola Lehmann, Jirka Lewandowski, Theres Matthieß, Nicolas Merz, Sven Regel, and Lisa Zehnter. 2022. Manifesto corpus.

Andrea Ceron, Luigi Curini, and Stefano M Iacus. 2019. *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Routledge, Abingdon, New York.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Rory Duthie and Katarzyna Budzynska. 2018. A deep modular rnn approach for ethos mining. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4041–4047. AAAI Press.

René D. Flores. 2017. Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *American Journal of Sociology*, 123(2):333–384.

R. Kelly Garrett, Shira Dvir Gvirsman, Benjamin K. Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal. 2014. Implications of pro- and counterattitudinal information exposure for affective polarization. *Human Communication Research*, 40(3):309–332.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Semi-supervised acquisition of Croatian sentiment lexicon. In: *International Conference on Text, Speech and Dialogue*, pages 166–173. Springer.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693, Valencia, Spain. Association for Computational Linguistics.

Hrvatski sabor. 2020. eDoc. http://edoc.sabor.hr/.

Swen Hutter, Edgar Grande, and Hanspeter Kriesi. 2016. *Politicising Europe: Integration and mass politics.* Cambridge University Press, Cambirdge.

[9] https://huggingface.co/classla/bcms-bertic-parlasent-bcs-ter

[10] https://huggingface.co/classla/bcms-bertic-parlasent-bcs-bi

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. 2019. The Origins and Consequences of Affective Polarization in the United States. *Annual Review of Political Science*, 22(1):129–146.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Filip Klubička, Gema Ramírez-Sánchez, and Nikola Ljubešić. 2016. Collaborative development of a rule-based machine translator between croatian and serbian. In: *Proceedings of the 19th Annual Conference of the Eu- ropean Association for Machine Translation*, pages 361– 367.

Ruud Koopmans and Paul Statham. 2006. Political Claims Analysis: Integrating Protest Event and Political Discourse Approaches. *Mobilization: An International Quarterly*, 4(2):203–221.

Taja Kuzman, Peter Rupnik, and Nikola Ljubesic. 2022. The ginco training dataset for web genre identification of documents out in the wild. *ArXiv*, abs/2201.03857.

Harold Dwight Lasswell. 1927. *Propaganda Technique in the World War*. Peter Smith, New York.

Dilin Liu and Lei Lei. 2018. The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context & Media*, 25:143–152.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy, August. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić – the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In: *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine, April. Association for Computational Linguistics.

Nikola Ljubešić. 2018. Word embeddings CLARIN.SI-embed.hr 1.0. Slovenian language resource repository CLARIN.SI.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Lilliana Mason. 2015. "I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization. *American Journal of Political Science*, 59(1):128–145.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2020. Talking War: Representation, Veterans and Ideology in Post-War Parliamentary Debates. *Government and Opposition*, 57(1):148–170.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2022a. CROCorp: Corpus of Parliamentary Debates in Croatia (v1.1.1). https://doi.org/10.5281/zenodo.6521372.

Michal Mochtak, Josip Glaurdić, and Christophe Lesschaeve. 2022b. SRBCorp: Corpus of Parliamentary Debates in Serbia (v1.1.1). https://doi.org/10.5281/zenodo.6521648.

Michal Mochtak, Josip Glaurdić, Christophe Lesschaeve, and Ensar Muharemović. 2022c. BiHCorp: Corpus of Parliamentary Debates in Bosnia and Herzegovina (v1.1.1). https://doi.org/10.5281/zenodo.6517697.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2022d. The sentiment corpus of parliamentary debates ParlaSent-BCS v1.0. Slovenian language resource repository CLARIN.SI.

Saif M. Mohammad. 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. https://arxiv.org/abs/2005.11882.

Nona Naderi and Graeme Hirst. 2016. Argumentation mining in parliamentary discourse. In: Matteo Baldoni, Cristina Baroglio, Floris Bex, Floriana Grasso, Nancy Green, Mohammad-Reza Namazi-Rad, Masayuki Numao, and Merlin Teodosia Suarez, editors, *Principles and Practice of Multi-Agent Systems*, pages 16–25, Cham. Springer.

Obinna Onyimadu, Keiichi Nakata, Tony Wilson, David Macken, and Kecheng Liu. 2013. Towards sentiment analysis on parliamentary debates in Hansard. In: *Revised Selected Papers of the Third Joint International Confer- ence on Semantic Technology – Volume 8388*, JIST 2013, page 48–50, Berlin, Heidelberg. Springer-Verlag.

Otvoreni Parlament. 2020. Početna. https://otvoreniparlament.rs/.

Parlamentarna skupština BiH. 2020. Sjednice. https://www.parlament.ba/?lang=bs.

Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, 44(1):97–131.

Christian Rauh. 2018. Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics*, 15(4):319–343.

Edward A. Shils and Morris Janowitz. 1948. Cohesion and Disintegration in the Wehrmacht in World War II. *Public Opinion Quarterly*, 12(2):315.

Juan M. Soler, Fernando Cuartero, and Manuel Roblizo. 2012. Twitter as a tool for predicting elections results. In: *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1194–1200.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney. Association for Computational Linguistics.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment.

Proceedings of the International AAAI Conference on Web and Social Media, 4(1).

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In: P. Sojka, I. Kopeček, K. Pala, and A. Horák, eds., *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Stephen M. Utych. 2018. Negative Affective Language in Politics. *American Politics Research*, 46(1):77–102.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.