State of the Art in Text Classification for South Slavic Languages: Fine-Tuning or Prompting?

Taja Kuzman Pungeršek[*], **Peter Rupnik**[*], **Ivan Porupski**[*], **Vuk Dinić**[*], **Nikola Ljubešić**[*†‡]
[*]Jožef Stefan Institute;
[†]Faculty of Computer and Information Science, University of Ljubljana;
[‡]Institute of Contemporary History;
Ljubljana, Slovenia
{taja.kuzman, peter.rupnik, ivan.porupski, vuk.dinic, nikola.ljubesic}@ijs.si

Abstract

Until recently, fine-tuned BERT-like models provided state-of-the-art performance on text classification tasks. With the rise of instruction-tuned decoder-only models, commonly known as large language models (LLMs), the field has increasingly moved toward zero-shot and few-shot prompting. However, the performance of LLMs on text classification, particularly on less-resourced languages, remains under-explored. In this paper, we evaluate the performance of current language models on text classification tasks across several South Slavic languages. We compare openly available fine-tuned BERT-like models with a selection of open-weight and closed-source LLMs across three tasks in three domains: sentiment classification in parliamentary speeches, topic classification in news articles and parliamentary speeches, and genre identification in web texts. Our results show that LLMs demonstrate strong zero-shot performance, often matching or surpassing fine-tuned BERT-like models. Moreover, when used in a zero-shot setup, LLMs perform comparably in South Slavic languages and English. However, we also point out key drawbacks of LLMs, including less predictable outputs, significantly slower inference, and higher computational costs. Due to these limitations, fine-tuned BERT-like models remain a more practical choice for large-scale automatic text annotation.

Keywords: LLM evaluation, text classification, large language models, South Slavic languages, sentiment identification, topic classification, genre identification

1. Introduction

Until recently, the dominant approach for text classification tasks relied on fine-tuning BERT-like transformer models on thousands of manually-annotated training examples. Recently, however, the field has shifted with the development of instruction-tuned decoder-only transformer models. These models, also commonly referred to as large language models (LLMs), which were originally developed primarily for text generation tasks, have demonstrated remarkable capabilities across a broad range of natural language processing (NLP) tasks, including text classification (Kuzman et al., 2023; Huang et al., 2023).

In this paper, we focus on South Slavic languages, where research on text classification tasks included in our study has, until recently, been limited or even non-existent (Kuzman and Ljubešić, 2023; Mochtak et al., 2024; Kuzman and Ljubešić, 2025). We take a first step toward systematically evaluating the current state of the art for text classification in these languages. Our evaluation is based on three text classification tasks in three different domains for which manually-annotated test datasets in South Slavic languages and fine-tuned BERT-like classifiers are freely available: sentiment classification of parliamentary speeches, topic classification in news articles, topic classification in parliamentary speeches, and automatic genre identification in web texts. These tasks span different domains and language styles, allowing for a comprehensive analysis of the performance of transformer-based models on text classification tasks. Specifically, we compare the performance of openly available fine-tuned BERT-like models with the zero-shot capabilities of both open-weight and closed-source LLMs used via prompting.

An important aspect of our study is to examine whether the performance of multilingual models on South Slavic languages is on par with their performance on English. This question is particularly relevant given that the evaluated large language models have been predominantly pretrained and instruction-tuned on English data.

By evaluating various models on a selection of text classification tasks in English and various South Slavic languages, we set out to test the following two hypotheses that are based on previous experiments with fine-tuned BERT-like models and LLMs on automatic genre identification (Kuzman et al., 2023), news topic classification (Kuzman and Ljubešić, 2025) and sentiment analysis in parliamentary texts (Mochtak et al., 2025):

H1 Zero-shot prompting with instruction-tuned large language models (LLMs) can achieve

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 1–17
11 May 2026.

Dataset	Lang	# Instances	# Labels	% Most and Least Frequent Label
Sentiment classification in parliamentary speeches				
ParlaSent-EN-test	EN	2600	3	40.8% (Neutral), 26.8% (Positive)
ParlaSent-HR-test	HR	1336	3	41.9% (Negative), 17.2% (Positive)
ParlaSent-SR-test	SR	1074	3	46.2% (Negative), 17.6% (Positive)
ParlaSent-BS-test	BS	190	3	47.9% (Negative), 14.7% (Positive)
Genre classification in web texts				
EN-GINCO	EN	272	8	23.5% (Information/Explanation), 0.4% (Legal)
X-GINCO-SL	SL	80	8	15% (Prose/Lyrical), 8.8% (Opinion/Argumentation)
X-GINCO-HR	HR	80	8	16.3% (Promotion), 7.5% (Instruction)
X-GINCO-MK	MK	80	8	15% (News), 1% (Opinion/Argumentation)
Topic classification in news articles				
IPTC-test-HR	HR	291	17	11.0% (Economy), 3.8% (Conflict, War and Peace)
IPTC-test-SL	SL	282	17	10.6% (Society), 3.2% (Conflict, War and Peace)
Topic classification in parliamentary speeches				
ParlaCAP-test-EN	EN	876	22	6.4% (Law and Crime), 2.1% (Culture)
ParlaCAP-test-HR	HR	869	22	8.5% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-SR	SR	874	22	7.1% (Government Operations), 1.7% (Immigration)
ParlaCAP-test-BS	BS	824	22	10.4% (Other), 0.5% (Culture)

Table 1: Information on test datasets in English (EN), Croatian (HR), Serbian (SR), Bosnian (BS), Slovenian (SL), and Macedonian (MK).

results comparable to the use of BERT-like models fine-tuned on training data that are similar to the test data.

H2 The performance of LLMs used in a zero-shot setup on text classification tasks on South Slavic test datasets is comparable to the performance on English test datasets.

2. Related Work

After the introduction of transformer architectures, BERT (bidirectional encoder representations from transformers) models have achieved state-of-the-art results in text classification tasks, outperforming earlier non-neural approaches, such as support vector machines (SVMs). They have also demonstrated strong cross-lingual zero-shot capabilities in various classification tasks, including automatic genre identification (Kuzman and Ljubešić, 2023), news topic classification (Petukhova and Fachada, 2023; De Clercq et al., 2020), and sentiment classification (Mochtak et al., 2024). However, these models still require fine-tuning on a training dataset, developed during manual annotation campaigns that are time-consuming and costly.

Instruction-tuned decoder-only transformer models, commonly referred to as large language models (LLMs), have recently shown strong performance in a range of classification tasks, even in zero-shot prompting setups that require no training data (Kuzman et al., 2023; Ljubešić et al., 2024a; Huang et al., 2023; Kuzman Pungeršek et al., 2026). They have achieved promising results on various natural language processing tasks, including stance detection (Zhang et al., 2022), implicit hate speech categorization (Huang et al., 2023), news topic classification (Kuzman and Ljubešić, 2025), automatic genre identification (Kuzman et al., 2023), causal commonsense reasoning (Ljubešić et al., 2024b), and machine translation (Hendy et al., 2023). Due to their promising performance, researchers have even started using them as data annotators, either by generating text and labels (Meng et al., 2022) or by annotating pre-existing texts (Kuzman and Ljubešić, 2025; Kuzman Pungeršek et al., 2026). Despite the growing interest in this topic, the majority of evaluations of LLMs used in

text classification tasks are limited only to English (Sun et al., 2023; Zhang et al., 2025; Kostina et al., 2025; Zhao et al., 2024). Systematic multilingual evaluations, especially which would include less-resourced languages such as those in the South Slavic group remain limited. Our work addresses this gap by providing a comparative evaluation of open-weight and closed-source LLMs with openly-available fine-tuned BERT-like models across four benchmark families comprising three diverse classification tasks and three different domains in South Slavic languages and English.

3. Benchmarks

The benchmarks (evaluation datasets) used in this study cover three text classification tasks, namely, sentiment identification, topic classification, and automatic genre identification, and three domains: parliamentary speeches, news articles and web texts. An overview of the datasets is provided in Table 1. The four benchmark families differ significantly in terms of language coverage, number of test instances, and label granularity.

The topic classification task is evaluated on two domains: 1) news articles, namely, the Croatian and Slovenian IPTC test datasets (Kuzman and Ljubešić, 2025), which comprise around 300 text instances per language, and 2) parliamentary speeches, namely, the Bosnian, Croatian, English and Serbian ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) that consist of approximately 820 to 880 instances per language. In the ParlaCAP benchmarks, an instance is a transcription of an utterance given by a parliamentary member in a parliamentary session.

The topic classification task involves the highest number of labels, that is, 17 news topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema[1] (IPTC, 2022), and 22 policy topic labels (21 major topics and a label *Other*) from the Comparative Agendas Project (CAP; Baumgartner et al., 2019) Master Codebook (Bevan, 2019).[2]

In contrast, the Bosnian, Croatian, English, and Serbian ParlaSent sentiment identification datasets (Mochtak et al., 2024; Mochtak et al., 2023) have a significantly lower granularity of labels, with only 3 categories. They are represented by the largest number of instances, ranging from 190 (Bosnian part) to 2600 (English part) sentence-level instances.

With 8 labels, the Croatian, English, Macedonian, and Slovenian GINCO genre datasets (Kuzman et al., 2023) represent a midpoint in label granularity among the four benchmark families. However, the genre identification task might be the most difficult one, as genre identification depends on the interpretation of full texts with the focus on author's purpose, the common function of the text, and the text's conventional form (Orlikowski and Yates, 1994). This complexity has also contributed to smaller test datasets in terms of the number of text instances, as manual annotation is more time-consuming. It is also important to note that, unlike the parliamentary datasets, the English portion of the genre datasets is not fully comparable to the South Slavic portions, which are label-balanced and contain fewer ambiguous instances. Nevertheless, the genre datasets remain valuable for evaluating model performance within each language.

All test datasets were manually annotated by annotators that are deemed reliable based on their satisfactory inter-annotator agreement, namely, Krippendorff's alpha (Krippendorff, 2018) values close to or above the 0.667 threshold for reliable annotation. To prevent large language models from incorporating the test datasets during their training phase, the test datasets are not publicly available, except for the ParlaSent benchmark family. Access to other datasets is granted on request from the corresponding authors. Further details on the test datasets are provided in Section A.1 of the Appendix.

4. Methodology

In this paper, we evaluate the main machine learning approaches that have recently been used for our selection of text classification tasks, with a focus on the comparison between the freely available fine-tuned BERT models and the open-weight and closed-source LLMs.[3] The models are evaluated on four families of test datasets that comprise South Slavic languages. The performance of the models is evaluated based on the micro-F1 and macro-F1 metrics, which enable assessment of the model performance at both the instance and label levels, respectively.

The following machine learning models are included in the evaluation:

- **dummy classifier**: a dummy classifier that predicts the most frequent class in the training data. To allow comparison, the dummy classifiers were trained on the same datasets that were used for fine-tuning the BERT-like models, mentioned below.

[1] https://show.newscodes.org/index.html?newscodes=medtop&lang=en-GB&startTo=Show

[2] https://www.comparativeagendas.net/pages/master-codebook

[3] The code for the model evaluation and analysis of results is available at https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic.

- **fine-tuned BERT-like classifiers**: in our study, we evaluate previously developed openly accessible multilingual fine-tuned BERT-like models that have been fine-tuned for the respective task, namely, the XLM-R-ParlaSent (Rupnik et al., 2023; Mochtak et al., 2024) model for sentiment identification in parliamentary texts, the X-GENRE classifier (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d, 2023) for automatic genre identification, the IPTC News Topic classifier (Kuzman and Ljubešić, 2025; Kuzman and Ljubešić, 2025) for news topic classification, and the ParlaCAP classifier (Kuzman Pungeršek et al., 2026; Kuzman Pungeršek and Ljubešić, 2025) for topic classification in parliamentary speeches. The XLM-R-ParlaSent and the ParlaCAP models are based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on parliamentary proceedings in 30 European languages (Mochtak et al., 2024). The XLM-R-ParlaSent model was fine-tuned on 13 thousand instances from the ParlaSent sentiment training dataset (Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian; Mochtak et al., 2024), while the ParlaCAP model was fine-tuned on the ParlaCAP-train dataset (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026) that comprises around 30 thousand speeches from parliamentary debates annotated with CAP topic labels, originating from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The X-GENRE classifier is based on the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a) in English and Slovenian; while the IPTC News Topic classifier is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) that was fine-tuned on the EMMediaTopic dataset (Kuzman and Ljubešić, 2024c) in Catalan, Croatian, Greek, and Slovenian. All fine-tuned models use the same classes as the test datasets used in our study.

- **open-weight and closed-source large language models**: we use closed-source OpenAI models, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`; OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`; OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`; OpenAI, 2025); a closed-source Gemini 2.5 Flash model (Comanici et al., 2025) by Google DeepMind; a closed-source Mistral Medium 3.1 model (`mistral-medium-2508`; Mistral AI, 2025) by Mistral AI; and four open-weight models, namely, the Meta LLaMA 3.3 model (Meta, 2024), the Gemma 3 model (Gemma Team et al., 2025), the Qwen 3 model (Yang et al., 2025), and the DeepSeek-R1-Distill model (`DeepSeek-R1-Distill-Qwen-14B`; Guo et al., 2025). It is important to note that while the LLaMA model was pretrained on a web text collection in various languages, it is said to support only 8 languages, namely English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai (Meta, 2024). The DeepSeek-R1-Distill model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that provides support for more than 29 languages – not including South Slavic languages though. In contrast, the Gemma 3 model is reported to support over 140 languages (Gemma Team et al., 2025), and the Qwen 3 model was pretrained on 119 languages (Yang et al., 2025). While closed-source models are said to be massively multilingual, with Gemini 2.5 models being pretrained on over 400 languages (Comanici et al., 2025), details on their language coverage are very limited.

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). OpenAI models were used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[4] that provides a unified API access to various closed-source models. To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. More details on the models and their implementation, including information on the availability of openly available models and fine-tuning datasets, are provided in Section A.2 of the Appendix.

All instruction-tuned LLMs are used in a zero-shot prompting setup, meaning that they receive only a task description and label definitions. All prompts and label definitions are written in English, while the instances are provided in the original languages (English or South Slavic). Changing the language of the prompt could introduce an additional factor that affects performance. In the presented experiments, our goal is to assess model performance on a specific task and language, rather than to evaluate their instruction-following abilities across different languages. The models are instructed to output a label, represented by a digit. The same prompt per benchmark family is used for all LLMs.

[4] `https://openrouter.ai/`

Prompts are provided in Figure 4 in Section A.2 of the Appendix.

5. Results

In this section, we evaluate the performance of the fine-tuned BERT-like models and the instruction-tuned LLMs on a selection of text classification tasks that include test datasets in South Slavic languages. First, in Section 5.1, we provide results on the four benchmark families with a focus on hypothesis H1, which expects that zero-shot prompting with LLMs can provide performance that is comparable to that of fine-tuned BERT-like models. In Section 5.2, we compare in more detail the performance of the closed-source and open-weight LLMs on the three text classification tasks, which is followed by a discussion on the advantages and limitations of LLMs for data annotation based on text classification tasks (Section 5.3). Lastly, in Section 5.4, we compare the performance of LLMs on English test datasets with their performance on South Slavic datasets, addressing hypothesis H2, which presumes that the available multilingual LLMs perform similarly on South Slavic languages as on English.

Model	Rank	Rank (EN)	Rank (South Slavic)
GPT-5	2.29	1.33	2.55
GPT-4o	2.36	2.00	2.45
Fine-Tuned BERT-Like Model	3.21	4.67	2.82
Gemini 2.5 Flash	3.50	3.33	3.55
Mistral Medium 3.1	5.36	5.00	5.45
Gemma 3	5.71	5.67	5.73
LLaMA 3.3	6.00	6.67	5.82
Qwen 3	7.43	7.00	7.55
GPT-3.5-Turbo	8.79	9.00	8.73
DeepSeek-R1-Distill	10.00	10.00	10.00

Table 2: Comparison of models based on their average rank (1 = best-performing, 10 = worst-performing) across all test datasets (first column), and averaged across English (second column) or South Slavic (third column) test datasets.

5.1. State of the Art in Text Classification Tasks

Figure 1 provides the results of model evaluation on our selection of text classification tasks. A consistent pattern emerges across all four benchmark families: LLMs, when used in a zero-shot prompting setup, achieve some of the highest scores. As shown in Table 2, which compares model rankings across tasks, LLMs achieve first place more often on average than the fine-tuned BERT-like model.

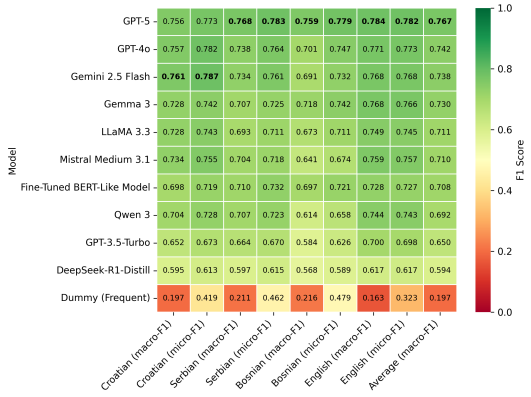
Figure 1a shows that both open-weight and closed-source LLMs, used in a zero-shot prompting setup on the sentiment identification task, achieve performance that is comparable or even significantly higher than that of a fine-tuned BERT-like model trained on a large manually-annotated sentiment dataset. The only models that consistently perform worse than the fine-tuned BERT-like model are GPT-3.5-Turbo and DeepSeek-R1-Distill. Sentiment classification appears broad enough that more potent LLMs can interpret label definitions effectively without task-specific fine-tuning, reducing the benefit of additional training.

In contrast, fine-tuned BERT-like models outperform most LLMs on automatic genre identification and topic classification tasks. These tasks depend on predefined label sets based on specific guidelines, and the strong performance of fine-tuned BERT-like models indicates that domain-specific fine-tuning on labelled data still offers an advantage over the general knowledge leveraged by LLMs in zero-shot setups. This advantage is particularly clear in genre identification for South Slavic texts, where the fine-tuned BERT-like model significantly outperforms LLMs. The likely reason for the fine-tuned model's very strong performance on South Slavic genre datasets is the curated nature of the test data – more challenging examples were removed before and during manual annotation, unlike in the English genre test dataset where the instances were randomly sampled from an English web corpus. Nevertheless, despite this limitation, the South Slavic test dataset remains valuable for comparing the performance of LLMs.

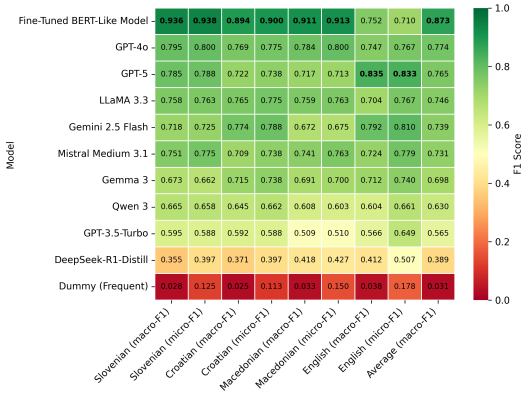
To conclude, since some LLMs used in a zero-shot prompting setup achieve higher or comparable results to fine-tuned BERT-like models across all classification tasks and languages, as shown in Table 2, we can confirm hypothesis H1, which proposed that zero-shot prompting with LLMs can perform comparably to fine-tuned BERT-like models.

5.2. Comparison of Large Language Models

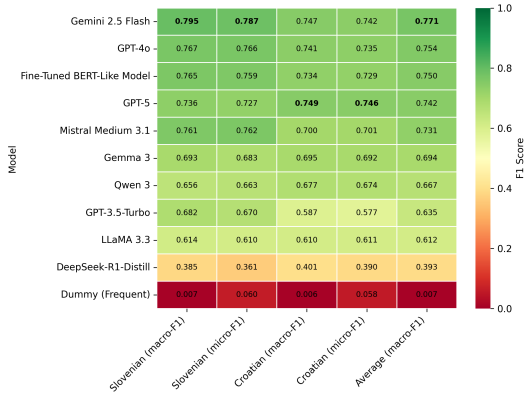
Figure 2 shows the performance of open-weight and closed-source LLMs, used via prompting, on the tasks of sentiment identification, automatic genre identification, news topic classification, and parliamentary topic classification. The DeepSeek-R1-Distill model is not included in the comparison, as it performs significantly worse than the other



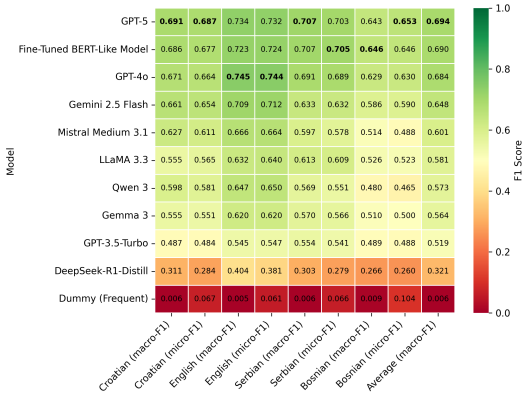
(a) Sentiment classification.



(b) Automatic genre identification.



(c) News topic classification.



(d) Parliamentary topic classification.

Figure 1: Micro-F1 and macro-F1 scores across models and languages on the test datasets for sentiment classification (Figure 1a), automatic genre identification (Figure 1b), and topic classification on news (Figure 1c) and parliamentary speeches (Figure 1d).

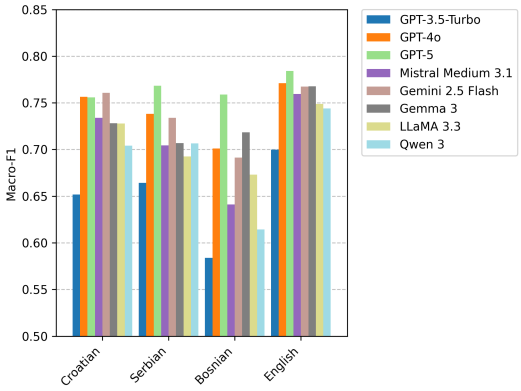
models, as shown in Figure 1.

While different models perform best across different languages and test datasets, a clear trend emerges: the top-performing models across all four benchmark families are the closed-source GPT-4o and GPT-5 from OpenAI, along with Gemini 2.5 Flash. Although GPT-5 is newer and reportedly more powerful, it does not outperform GPT-4o on all benchmarks. Among open-weight models, Gemma 3 generally achieves the best results in sentiment identification (Figure 2a) and news topic classification (Figure 2c). For automatic genre identification (Figure 2b) and parliamentary topic classification (Figure 2d), rankings of open-weight models vary by language. Overall, the weakest performance is observed with the older closed-source GPT-3.5-Turbo model, highlighting the rapid progress in both open-weight and closed-source model development.

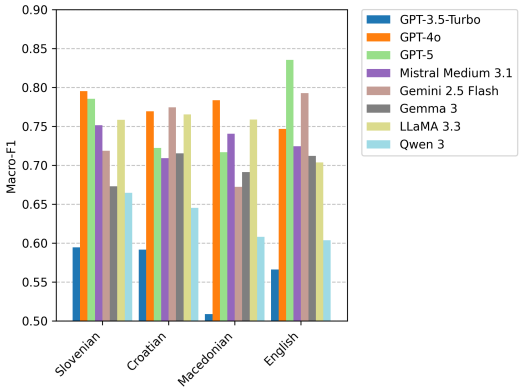
5.3. Advantages and Disadvantages of LLMs

A clear advantage of LLMs is that they do not require manually-annotated training data for specific tasks, yet still achieve strong performance when provided only with task instructions and brief label descriptions. However, these models are significantly more computationally expensive than fine-tuned BERT-like models. While closed-source models deliver the best performance, as shown in previous sections, they come with several limitations: they are costly to use, their architectures and pre-training data are not publicly disclosed, and access through APIs hinders reproducibility, in contrast to open-weight LLMs and fine-tuned BERT-like models.

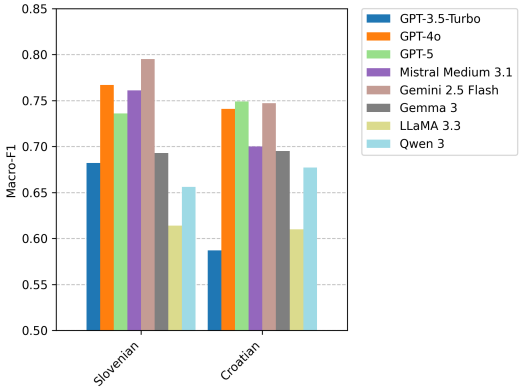
What is more, the inference speed of all LLMs is significantly slower than that of a fine-tuned BERT-like model. As shown in Figure 3, the fine-tuned BERT-like model achieves one of the highest macro-F1 scores on the topic classification task for parliamentary speeches, while maintaining a very low inference time of just 0.02 seconds per



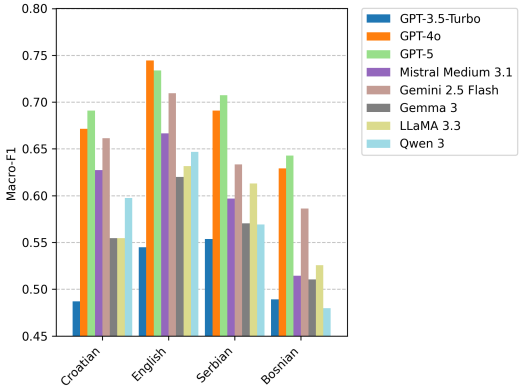
(a) Sentiment classification.



(b) Automatic genre identification.



(c) News topic classification.



(d) Parliamentary topic classification.

Figure 2: Comparison of LLMs used in a zero-shot prompting setup on sentiment identification (Figure 2a), automatic genre identification (Figure 2b), and topic classification on news (Figure 2c) and parliamentary speeches (Figure 2d).

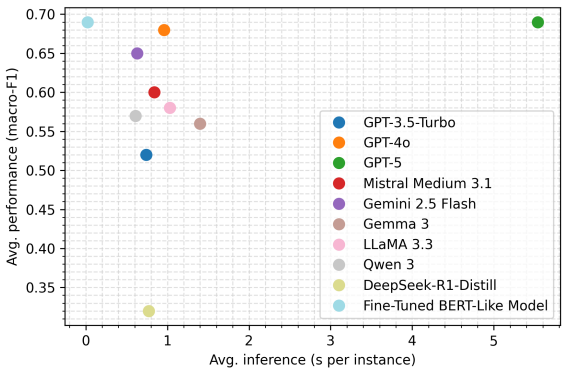


Figure 3: Comparison of models on the parliamentary topic classification based on their inference speed (seconds per instance) and performance (macro-F1 scores), both averaged across all four languages.

instance. In contrast, most LLMs have inference times between 0.6 and 1.4 seconds per instance, making them three to seven times slower for annotating the same dataset. The slowest model, GPT-5, takes 5.5 seconds per instance, which renders it impractical for large-scale automatic annotation of text collections. In this regard, fine-tuned BERT-like models offer a key advantage due to their lower computational cost and higher inference speed. Moreover, they can be trained on training data that is annotated by LLMs using the recently introduced LLM teacher-student paradigm (Kuzman and Ljubešić, 2025), which considerably reduces the effort needed to develop task-specific models.

Another limitation of LLMs, as revealed by the experiments, is their occasional deviation from the defined label set. This issue was especially noticeable in topic classification and, to a lesser extent, in genre identification. The highest rate of label hallucination was found in the DeepSeek-R1-Distill model, which produced non-existing labels for 8% of instances in the news topic test datasets and 4% in the genre test dataset. Similar issues were also observed, though much less frequently (less than 1%), with the LLaMA 3.3, Gemma 3, Qwen 3 and Mistral Medium 3.1 models. In contrast, fine-tuned

BERT-like models do not suffer from this issue, as they output probabilities for the predefined classes.

Model	Difference (sentiment)	Difference (topic)
GPT-5	0.02	0.05
GPT-4o	0.04	0.08
Gemini 2.5 Flash	0.04	0.08
Gemma 3	0.05	0.07
LLaMA 3.3	0.05	0.07
Mistral Medium 3.1	0.07	0.09
Qwen 3	0.07	0.10
GPT-3.5-Turbo	0.07	0.03

Table 3: Difference between model performance in macro-F1 scores obtained on sentiment and topic classification in parliamentary texts on English versus the average macro-F1 scores on South Slavic languages.

5.4. Performance on English versus on South Slavic languages

The sentiment identification ParlaSent and the topic classification ParlaCAP benchmark families comprise test datasets in South Slavic languages and English that were constructed with the same methodology. Thus, they also allow for a comparison of the performance of the LLMs on English, a highly resourced language, with South Slavic languages, which are significantly less represented in the pretraining and instruction-tuning datasets used to develop large language models.

As shown in Table 3, the differences in macro-F1 scores between English and the average of macro-F1 scores for South Slavic languages are relatively small for sentiment identification, ranging from 2 to 7 points. For topic classification, the performance gap is slightly larger, ranging from 3 to 10 points. This is likely due to the increased difficulty of the task, which involves greater label granularity: 22 labels compared to just 3 in sentiment classification. These findings partially confirm hypothesis H2, which stated that LLMs, when used in a zero-shot setup, perform comparably on text classification tasks in South Slavic languages as they do on English.

Interestingly, even the open-weight LLaMA 3.3 model – reported to support only eight languages, excluding the South Slavic group – does not show a substantial performance drop when applied to South Slavic languages compared to English.

6. Conclusion

In this paper, we evaluated how well current machine learning technologies handle text classification tasks in South Slavic languages. We compared fine-tuned BERT-like models with decoder-only generative large language models (LLMs) that are used in a zero-shot prompting setup across three tasks and three text domains: sentiment classification in parliamentary texts, news topic classification, topic classification in parliamentary texts, and automatic genre identification on web texts.

Our results show that LLMs used with prompting, where only a brief task description and labels were provided, achieved strong results across all tasks and languages, particularly the closed-source GPT-4o (OpenAI, 2024), GPT-5 (OpenAI, 2025) and Gemini 2.5 Flash (Comanici et al., 2025) models. The performance of LLMs is comparable or higher than that of fine-tuned BERT-like models specialized for the tasks. On the sentiment identification task, most open-weight and closed-source LLMs outperformed the fine-tuned model, demonstrating strong general knowledge of the notion of sentiment. For genre and topic classification, however, fine-tuning BERT-like models remains beneficial, as these tasks rely on predefined label sets and fine-tuning aligns the models more closely with the task requirements.

Interestingly, LLMs perform similarly in English and South Slavic languages, with rather minor drops in micro- and macro-F1 scores, namely a drop of 2 to 7 points in terms of macro-F1 scores on sentiment classification, and a slightly higher drop from 3 to 10 points on topic classification in parliamentary texts. This suggests that the gap in multilingual performance is smaller than expected, even for open-weight models not explicitly dedicated to these languages.

Although large language models offer impressive zero-shot performance and reduce the need for annotated data, they come with higher computational costs and are more prone to producing invalid labels. Moreover, their inference speed is at least three times slower than that of the fine-tuned BERT-like models. Thus, their use in use cases with extensive data to be processed, such as automatic enrichment of large corpora with text categories, remains impractical due to their high computational demands. In contrast, fine-tuned BERT-like models are more computationally efficient and can be better tailored to the specific characteristics of a task and its domain. They remain a practical and reliable choice for text classification tasks, especially when computational resources are limited, high inference speed is desired or output reliability is critical. Moreover, it is possible to combine the strengths of both approaches, as proposed by

the LLM teacher-student paradigm (Kuzman and Ljubešić, 2025): LLMs can be used to automatically annotate training data, reducing the need for costly and time-consuming manual annotation, while fine-tuned BERT-like models can then be trained on these datasets.

This study represents only an initial step to systematically benchmark text classification performance in South Slavic languages. Although our evaluation includes four diverse benchmark families, some of the test datasets remain relatively small. Future work will aim to increase dataset sizes, include more South Slavic languages and dialects, and introduce additional classification tasks. As new large language models continue to emerge rapidly, it will also be important to establish ongoing evaluations to track whether their performance continues to improve, particularly on South Slavic languages. Importantly, this study only evaluated the performance of LLMs in a zero-shot prompting setup. In future work, we plan to extend the evaluation to include few-shot prompting and fine-tuning on training data. To support further research and facilitate reproducibility, we have made all code, evaluation scripts, and results publicly available.[5] Additionally, we have developed an interactive dashboard[6] that enables users to explore the results of our evaluation of large language models on the tasks presented in this paper, as well as on additional commonsense reasoning tasks. The dashboard is an ongoing project that monitors the performance of newly released large language models on South Slavic languages and dialects. In future work, we plan to expand both the range of tasks and the coverage of South Slavic languages and dialects included in the dashboard.

7. Ethical Considerations and Limitations

Our study has several limitations that should be acknowledged. First, while we aimed to include a broad set of South Slavic languages, some – most notably Bulgarian – were not covered in our experiments. We assume that the performance on Bulgarian would be similar to that observed for Macedonian, given their close linguistic proximity, or the results for Bulgarian could be slightly better, as Macedonian is comparatively more low-resourced. Moreover, due to the high computational cost of evaluating the LLMs on all the test datasets and the financial cost associated with the use of closed-source models, each model was evaluated on each dataset only once. This setup prevents us from fully estimating the variance of the results, however, based on our preliminary experiments, we expect this variance to be relatively small. Finally, the scope of our evaluation remains limited in terms of test datasets, language coverage and tasks. Expanding the range of benchmarks would allow for a more comprehensive validation of our findings, particularly regarding the hypothesis that LLMs can perform on par with fine-tuned BERT-like models across diverse natural language understanding tasks, languages and language varieties.

[5] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

[6] See the CLASSLA LLM Evaluation Dashboard for South Slavic Languages at https://www.clarin.si/classla-llm-dashboard/.

8. Acknowledgements

We would like to thank the developers of the llm.ijs.si service (Marić et al., 2025) for establishing the LLM inference platform deployed at the Jožef Stefan Institute, which provided convenient access to the open-weight large language models used in this study. We also thank the annotators of the test datasets for their diligence and the time devoted to manual annotation, which resulted in the high-quality evaluation datasets used in this work. Lastly, we would like to thank the CLASSLA knowledge centre for South Slavic languages and the Slovenian CLARIN.SI infrastructure for their valuable support.

This work was supported in part by the projects "Spoken Language Resources and Speech Technologies for the Slovenian Language" (Grant J7-4642), "Large Language Models for Digital Humanities" (Grant GC-0002), the research programme "Language Resources and Technologies for Slovene" (Grant P6-0411), all funded by the ARIS Slovenian Research and Innovation Agency, and the research project "Embeddings-based techniques for Media Monitoring Applications" (L2-50070), co-funded by the Kliping d.o.o. agency. The authors acknowledge the OSCARS project – and its ParlaCAP cascading grant project –, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470. This research is co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Marta Bañón, Miquel Esplà-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *23rd Annual Conference of the European Association for Machine Translation*, pages 301–302.

Frank R Baumgartner, Christian Breunig, and Emiliano Grossman. 2019. *Comparative Policy Agendas: Theory, Tools, Data*. Oxford University Press.

Shaun Bevan. 2019. Gone Fishing. *Comparative Policy Agendas: Theory, Tools, Data*, pages 17–34.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Orphée De Clercq, Luna De Bruyne, and Véronique Hoste. 2020. News topic classification as a first step towards diverse news recommendation. *Computational Linguistics in the Netherlands Journal*, 10:37–55.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2025. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*, 59(3):2071–2102.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

IPTC. 2022. Groups of NewsCodes. https://iptc.org/standards/newscodes/groups/#descrncd. Accessed October 29, 2024.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The TenTen corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Lancaster University.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

Arina Kostina, Marios D Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *arXiv preprint arXiv:2501.08457*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage Publications.

Taja Kuzman and Nikola Ljubešić. 2023. Automatic genre identification: a survey. *Language Resources and Evaluation*, pages 1–34.

Taja Kuzman and Nikola Ljubešić. 2025. LLM Teacher-Student Framework for Text Classification With No Manually Annotated Data: A Case Study in IPTC News Topic Classification. *IEEE Access*, 13:35621–35633.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The GINCO Training Dataset for Web Genre Identification of Documents Out in the Wild. In *Language Resources and Evaluation Conference*, pages 1584–1594, Marseille, France. European Language Resources Association.

Taja Kuzman Pungeršek, Peter Rupnik, Daniela Širinić, and Nikola Ljubešić. 2026. Supercharging Agenda Setting Research: The ParlaCAP Dataset of 28 European Parliaments and a Scalable Multilingual LLM-Based Classification. *arXiv preprint arXiv:2602.16516*.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024a. DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Ivan Vulić, Fabian Schmidt, and Goran Glavaš. 2024b. JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 209–219.

Nikola Marić, Boshko Koloski, Damjan Demšar, Jan Jona Javoršek, and Sašo Džeroski. 2025. Running large language models locally: design and operational insights with llm.ijs.si. *International conference AI for science 2025: Ljubljana, Slovenia, 22.09.2025-26.09.2025*, page 77.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Meta. 2024. Llama 3.3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: June 26, 2025.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv preprint arXiv:2004.03705*.

Mistral AI. 2025. Medium is the new large. https://mistral.ai/news/mistral-medium-3. Accessed: October 10, 2025.

Michal Mochtak, Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2025. Parlasent: mapping sentiment in political discourse with large language models. *Political Research Exchange*, 7(1):2508377.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16024–16036.

OpenAI. 2023. ChatGPT General FAQ. https://help.openai.com/en/articles/6783457-chatgpt-general-faq. Accessed: June 26, 2025.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed September 11, 2024.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/. Accessed: October 10, 2025.

Wanda J Orlikowski and JoAnne Yates. 1994. Genre repertoire: The structuring of communicative practices in organizations. *Administrative science quarterly*, pages 541–574.

Alina Petukhova and Nuno Fachada. 2023. MN-DS: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5):74.

Qwen Team. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen Team. 2024b. Qwen2.5: A party of foundation models. Accessed: June 26, 2025.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. 2025. Pushing the limit of LLM capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528.

Hang Zhao, Qile P Chen, Yijing Barry Zhang, and Gang Yang. 2024. Advancing Single and Multi-task Text Classification through Large Language Model Fine-tuning. *arXiv preprint arXiv:2412.08587*.

10. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and others. 2024. *Multilingual comparable corpora of parliamentary debates ParlaMint 4.1*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1912.

Kuzman, Taja and Ljubešić, Nikola. 2023. *Multilingual text genre classification model X-GENRE*. Hugging Face. PID https://doi.org/10.57967/hf/0927.

Kuzman, Taja and Ljubešić, Nikola. 2024a. *English-Slovenian text genre dataset X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1960.

Kuzman, Taja and Ljubešić, Nikola. 2024b. *Genre-enriched web corpora MaCoCu-Genre*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1969.

Kuzman, Taja and Ljubešić, Nikola. 2024c. *Multilingual IPTC Media Topic dataset EMMediaTopic 1.0*. Slovenian Language Resource Repository CLARIN.SI. PID http://hdl.handle.net/11356/1991.

Kuzman, Taja and Ljubešić, Nikola. 2024d. *Multilingual text genre classification model X-GENRE*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1961.

Kuzman, Taja and Ljubešić, Nikola. 2025. *Multilingual IPTC News Topic Classifier*. Hugging Face. PID https://doi.org/10.57967/hf/4709.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2025. *Multilingual ParlaCAP model for CAP Topic Classification in Parliamentary Speeches*. Hugging Face. PID https://doi.org/10.57967/hf/6684.

Kuzman Pungeršek, Taja and Ljubešić, Nikola. 2026. *Multilingual training dataset for CAP policy topic classification ParlaCAP-train*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/2093.

Ljubešić, Nikola and Rupnik, Peter and van Noord, Rik. 2023. *Multilingual parliamentary model XLM-R-parla*. Hugging Face. PID https://doi.org/10.57967/hf/6717.

Mochtak, Michal and Rupnik, Peter and Meden, Katja and Ljubešić, Nikola. 2023. *The multilingual sentiment dataset of parliamentary debates ParlaSent 1.0*. Slovenian language resource repository CLARIN.SI. PID http://hdl.handle.net/11356/1868.

Rupnik, Peter and Ljubešić, Nikola and Mochtak, Michal. 2023. *Multilingual parliament sentiment regression model XLM-R-ParlaSent*. Hugging Face. PID https://doi.org/10.57967/hf/6718.

A. Appendix

A.1. Benchmarking Datasets

In this section, we provide additional information on the datasets used for benchmarking the models on sentiment identification, topic classification, and genre identification tasks in this study.

ParlaSent test datasets for sentiment classification in parliamentary speeches include Croatian, Serbian, Bosnian, and English data from the multilingual sentiment dataset of parliamentary debates ParlaSent 1.0 (Mochtak et al., 2024, Mochtak et al., 2023).[7] The dataset comprises sentences that were randomly sampled from Croatian, Serbian, Bosnian and British parliamentary corpora and manually annotated with reported inter-annotator agreement ranging from 0.53 to 0.66 in Krippendorff's alpha (Krippendorff, 2018). The annotation involved a more granular six-level sentiment polarity scale that has been mapped to a three-level sentiment polarity scale which we use in our experiments: negative (0), neutral (1), and positive (2).

GINCO datasets for automatic genre identification comprise the English EN-GINCO dataset (Kuzman et al., 2023) and a multilingual X-GINCO dataset from the AGILE benchmark for Automatic Genre Identification.[8] The test instances were sampled from the enTenTen20 English web corpus (Jakubíček et al., 2013) and the MaCoCu multilingual web corpus collection (Bañón et al., 2022). They were manually annotated by experts with a background in linguistics and computational linguistics who had experience with previous genre annotation campaigns (Kuzman et al., 2022, 2023) where they reached an acceptable inter-annotator agreement of 0.71 in nominal Krippendorff's alpha (Krippendorff, 2018). While the X-GINCO dataset comprises numerous European languages, for the purposes of this study, we focus on three South Slavic languages: Croatian, Macedonian, and Slovenian. The test datasets use the X-GENRE annotation schema (Kuzman et al., 2023) that includes the following genre labels: *Information/Explanation*, *News*, *Instruction*, *Opinion/Argumentation*, *Forum*, *Prose/Lyrical*, *Legal* and *Promotion*. While EN-GINCO and X-GINCO datasets have been annotated by the same annotator with the same schema, one should note that there are important differences between them in terms of their construction – the English test dataset was sampled randomly from the web corpus, resulting in an unbalanced label distribution, while the X-GINCO datasets were curated with more deliberate interventions to ensure a balanced label distribution and a more controlled sampling process. Consequently, the X-GINCO datasets comprise fewer ambiguous instances and could be regarded as an easier test dataset.

IPTC News Topic test datasets (Kuzman and Ljubešić, 2025) comprise Croatian and Slovenian news articles extracted from the MaCoCu-Genre web corpus collection (Kuzman and Ljubešić, 2024b) and manually annotated by one annotator. The reliability of the annotator was confirmed on a sample of data that was annotated by an additional annotator. The two annotators reached an acceptable inter-annotator agreement of 0.73 in nominal Krippendorff's alpha (Krippendorff, 2018). Text instances are annotated with 17 topic labels from the top level of the IPTC NewsCodes Media Topic hierarchical schema, developed by the International Press Telecommunications Council (IPTC) (IPTC, 2022). The datasets are more or less balanced by labels.

ParlaCAP test datasets (Kuzman Pungeršek et al., 2026) comprise parliamentary speeches in Bosnian, Croatian, English, and Serbian, sourced from the ParlaMint 4.1 dataset (Erjavec et al., 2024; Erjavec et al., 2025). These speeches were annotated by a single expert annotator using the 21 CAP categories from the official CAP schema (Baumgartner et al., 2019), along with an additional *Other* label. The datasets are approximately balanced across labels. To assess the annotation quality, the Croatian dataset was independently annotated by two additional annotators. Inter-annotator agreement between the expert annotator and the others ranged from 0.62 to 0.68 in Krippendorff's alpha, which is around the threshold of 0.67 typically considered acceptable for annotation reliability (Krippendorff, 2018).

A.2. Models

In the following subsections, we outline the models included in the evaluation – the fine-tuned BERT-like classifiers (Section A.2.1) and the open-weight and closed-source LLMs (Section A.2.2).

A.2.1. Fine-Tuned BERT-like Models

BERT (bidirectional encoder representations from transformers) deep neural models (Kenton and Toutanova, 2019) have revolutionized the field of

[7] Available in the CLARIN.SI repository at http://hdl.handle.net/11356/1585 and in the Hugging Face repository at https://huggingface.co/datasets/classla/ParlaSent.

[8] https://github.com/TajaKuzman/AGILE-Automatic-Genre-Identification-Benchmark

natural language processing (NLP), outperforming non-neural methods across various NLP tasks. They have a more complex and computationally expensive architecture featuring transformers – neural networks with self-attention mechanisms (Vaswani et al., 2017) – that significantly improves the efficiency of training the models on massive text data. Similarly to decoder-only transformer models, BERT models are pretrained on massive amounts of texts, possibly in multiple languages, which establishes their ability to encode the words and texts in high-dimensional vector spaces (Minaee et al., 2020) and enables their application even across languages in a zero-shot classification scenario. To develop BERT-based classifiers, the pretrained models are trained, that is, fine-tuned, on a training dataset comprising text instances annotated with labels. In our study, we evaluate openly-accessible multilingual fine-tuned BERT-like models that have already been developed in recent related research. Namely, we evaluate the following models:

- **IPTC News Topic classifier**[9] (Kuzman and Ljubešić, 2025) is a multilingual fine-tuned BERT-like model for news topic classification according to the top-level IPTC NewsCodes schema (IPTC, 2022). The model is based on the large-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on 15,000 training text instances from the EMMediaTopic[10] dataset (Kuzman and Ljubešić, 2024c). The training dataset contains news article instances in four languages: Catalan, Croatian, Greek, and Slovenian. The training dataset was annotated using an LLM that was shown to achieve annotation reliability comparable to that of human annotators (Kuzman and Ljubešić, 2025). This approach is based on the novel methodology that uses the LLM teacher-student pipeline to develop BERT-like classifiers in the absence of manually-annotated training data.
- **XLM-R-ParlaSent** (Rupnik et al., 2023; Mochtak et al., 2024) is a domain-specific multilingual transformer model for sentiment identification in parliamentary texts. It is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023) that was developed by additionally pretraining the large-sized XLM-RoBERTa model (Conneau et al., 2020) on 1.72 billion words from parliamentary proceedings in 30 European languages. To develop the XLM-R-ParlaSent model,[11] the pretrained XLM-R-Parla model was fine-tuned on the ParlaSent sentiment training dataset (Mochtak et al., 2024; Mochtak et al., 2023) in seven European languages (Bosnian, Croatian, Czech, English, Serbian, Slovak, and Slovenian). The training dataset[12] comprises 13,000 instances sampled from parliamentary proceedings and manually annotated with sentiment labels.
- **ParlaCAP classifier**[13] (Kuzman Pungeršek and Ljubešić, 2025; Kuzman Pungeršek et al., 2026) is a domain-specific multilingual transformer model for topic classification in parliamentary texts based on the CAP schema (Baumgartner et al., 2019). As the XLM-R-ParlaSent model, this model is based on the XLM-R-parla pretrained model (Ljubešić et al., 2023; Mochtak et al., 2024). The XLM-R-parla model was then fine-tuned on the ParlaCAP-train dataset[14] (Kuzman Pungeršek and Ljubešić, 2026; Kuzman Pungeršek et al., 2026). The training dataset comprises around 30 thousand speeches from parliamentary debates from the ParlaMint 4.1 parliamentary datasets (Erjavec et al., 2024; Erjavec et al., 2025) in 29 European languages. The training dataset was annotated with the CAP categories by a GPT-4o (OpenAI, 2024) model used in a zero-shot prompting setup, following the LLM teacher-student framework (Kuzman and Ljubešić, 2025). Based on the inter-annotator agreement, calculated on a sample that was annotated by three human annotators and the LLM annotator, the agreement between the LLM and the human annotators was comparable to the agreement between the human annotators themselves. This indicates that the LLM performs as reliably as human annotators on this task, supporting its use for annotating the training data.
- **X-GENRE classifier** (Kuzman et al., 2023; Kuzman and Ljubešić, 2024d) is a multilingual fine-tuned BERT-like model for automatic genre identification.[15] The model is based on

[9]The IPTC News Topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/multilingual-IPTC-news-topic-classifier.

[10]The EMMediaTopic training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/1991.

[11]The XLM-R-ParlaSent model is accessible in the Hugging Face repository at https://huggingface.co/classla/xlm-r-parlasent.

[12]The ParlaSent training and test datasets are freely available in the CLARIN.SI repository at http://hdl.handle.net/11356/1868.

[13]The ParlaCAP topic classifier is available in the Hugging Face repository at https://huggingface.co/classla/ParlaCAP-Topic-Classifier.

[14]The ParlaCAP-train training dataset is available in the CLARIN.SI repository at http://hdl.handle.net/11356/2093.

[15]The X-GENRE classifier is freely available in the

the base-sized XLM-RoBERTa model (Conneau et al., 2020) and was fine-tuned on the training split of the X-GENRE dataset (Kuzman and Ljubešić, 2024a), which contains 1,772 text instances in Slovenian and English, manually-annotated with genre labels from the X-GENRE schema (Kuzman et al., 2023).

A.2.2. Instruction-Tuned Large Language Models

As the BERT models, decoder-only large language models are based on a transformer deep neural architecture and are pretrained on massive text collections. However, while the development of fine-tuned BERT-like classifiers necessitates large amounts of annotated training data, recent advances in the field have shown that the instruction-tuned LLMs are capable of text classification in a zero-shot or few-shot prompting setups which do not require any training data. We assess the performance of the following large language models:

- **OpenAI models**, namely the GPT-3.5-Turbo (`gpt-3.5-turbo-0125`) (OpenAI, 2023), GPT-4o (`gpt-4o-2024-08-06`) (OpenAI, 2024) and the GPT-5 (`gpt-5-2025-08-07`) (OpenAI, 2025). These closed-source instruction-tuned LLMs were developed by OpenAI. OpenAI states that the models are trained on large multilingual web corpora, however, specific details about the training data, procedures, and architecture are not publicly known.
- **Gemini 2.5 Flash model** (Comanici et al., 2025) is a closed-source multilingual and multimodal instruction-tuned LLM by Google DeepMind. The model is reported to be pretrained on over 400 languages (Comanici et al., 2025), however, details on the language coverage are not available.
- **Mistral Medium 3.1 model** (`mistral-medium-2508`) (Mistral AI, 2025) is a closed-source multimodal instruction-tuned model by Mistral AI. Available details on the model architecture and language coverage are very limited.
- **LLaMA 3.3 model**[16] (Meta, 2024) is an open-weight instruction-tuned multilingual LLM, developed by Meta, with 70 billion parameters. The model was pretrained on a web text collection in various languages, however, it is reported to support only 8 languages, namely, English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.
- **Gemma 3 model**[17] (Gemma Team et al., 2025) is an open-weight multilingual instruction-tuned LLM, developed by Google DeepMind. The model was pretrained on multimodal data with large quantities of multilingual texts and is reported to support over 140 languages. We use the model in 27 billion parameter size.
- **DeepSeek-R1-Distill**[18] (Guo et al., 2025) is an open-weight reasoning LLM, developed by DeepSeek AI. We use the distilled model in 14 billion parameter size, namely the `DeepSeek-R1-Distill-Qwen-14B` model. The model is based on the Qwen 2.5 model (Qwen Team, 2024b,a) that was fine-tuned using a dataset curated with the DeepSeek-R1 reasoning model. The Qwen 2.5 model provides multilingual support for over 29 languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, and Arabic.
- **Qwen 3**[19] (`Qwen3-2504`) (Yang et al., 2025) is an open-weight LLM, developed by Alibaba Cloud. We use the model with the 32 billion parameter size, namely, the `qwen3:32b` model. The model is said to support over 100 languages and dialects (Yang et al., 2025).

Open-weight models were installed locally and executed via the Ollama API service (Marić et al., 2025). We use the quantized versions of the models as they are available through the Ollama library.[20] OpenAI models are used through the chat completion endpoint via the OpenAI API, whereas other closed-source models were accessed through the OpenRouter platform[21] that provides a unified API access to various closed-source models.

To prevent any bias, all models were used with their default parameters. The only parameter that we defined is the temperature which we set to 0 to ensure a more deterministic behaviour of the models. The same prompts were used for all open-weight and closed-source models. In Figure 4, we provide prompts that were provided to the LLMs for zero-shot text classification, namely for sentiment classification (Figure 4a), automatic genre identification (Figure 4b), news topic classification (Figure 4c) and topic classification in parliamentary

Hugging Face repository at https://doi.org/10.57967/hf/0927 and the CLARIN.SI repository at http://hdl.handle.net/11356/1961.

[16] https://ollama.com/library/llama3.3

[17] https://ollama.com/library/gemma3

[18] https://ollama.com/library/deepseek-r1:14b

[19] https://ollama.com/library/qwen3

[20] https://ollama.com/library

[21] https://openrouter.ai/

speeches (Figure 4d). For more details on the setups used to apply fine-tuned BERT-like models and instruction-tuned LLMs to the test datasets, refer to the code published on GitHub.[22]

[22] https://github.com/TajaKuzman/Benchmarking-Text-Classification-on-South-Slavic

```
### Task
    Your task is to classify the provided parliamentary text into a sentiment
label, meaning that you need to recognize whether the speaker's sentiment towards the topic
is negative, neutral, positive or somewhere in between. You will be provided with an excerpt
from a parliamentary speech in {lang} language, delimited by single quotation marks. Always
provide a label, even if you are not sure.


### Output format
    Return a valid JSON dictionary with the following key: 'sentiment' and a
value should be an integer which represents one of the labels according to the following
dictionary: {sentiment_description}.

    Text: '{text}'
```

(a) Sentiment classification.

```
### Task
    Your task is to classify the following text according to genre. Genres are
text types, defined by the function of the text, author's purpose and form of the text.
Always provide a label, even if you are not sure.

### Output format
    Return a valid JSON dictionary with the following key: 'genre' and a
value should be an integer which represents one of the labels according to the following
dictionary: {labels_dict}.


    Text: '{text}'
```

(b) Automatic genre identification.

```
### Task
    Your task is to classify the provided text into a topic label, meaning that you
need to recognize what is the topic of the text. You will be provided with a news text,
delimited by single quotation marks. Always provide a label, even if you are not sure.

### Output format
    Return a valid JSON dictionary with the following key: 'topic' and a value
should be an integer which represents one of the labels according to the following
dictionary: {label_dict_with_description}.


    Text: '{text}'
```

(c) News topic classification.

```
### Task
    Your task is to classify the provided text into a policy agenda topic label, meaning that you need to recognize what
is the predominant topic of the text. You will be provided with an excerpt from a parliamentary speech from the {parl} parliament in
{language} language, delimited by single quotation marks. Always provide a label, even if you are not sure.


    Follow the following rule: if the speech mentions a policy area and a policy instrument (e.g., taxes, laws), pick
the label based on the area, not the instrument (e.g., annotate mortgage tax changes with 14 (Housing), law on education with 6
(Education)).


### Output format
    Return a valid JSON dictionary with the following key: 'topic' and a value should be an integer which represents one
of the labels according to the following dictionary: {majortopics_description}.


    Text: '{text}
```

(d) Parliamentary topic classification.

Figure 4: The prompts that are provided to the LLMs for the sentiment identification task (Figure 4a), automatic genre identification (Figure 4b), and topic classification on news (Figure 4c) and parliamentary speeches (Figure 4d). The prompts comprise the description of the task and labels with short descriptions.