Intent classification for Slavic languages

Jan Busz
University of Warsaw
jb405986@students.mimuw.edu.pl

Michał Orzyłowski
University of Warsaw
mo418334@students.mimuw.edu.pl

Błażej Pałkus
University of Warsaw
bp385954@students.mimuw.edu.pl

Tomasz Siłkowski
University of Warsaw
ts407106@students.mimuw.edu.pl

Jan Ludziejewski
Supervisor

Abstract

Intent classification is a relevant area of research in machine learning, with applications in domains such as marketing, product design and dialogue systems. This paper presents a method for intent classification of utterances in Slavic languages. It leverages pre-trained language models such as HerBERT, XLM and mT5 as foundational frameworks. The approach involves fine-tuning these backbones along with classification heads and then training them on either language-specific or combined utterances in Polish, Slovak and Russian available in MASSIVE dataset.

1 Introduction

As we delve further into the era of digital transformation and automation, mastering the understanding of natural language emerges as a crucial progression in artificial intelligence (AI). At the core of this advancement lies intent classification, a fundamental component of conversational AI systems. It facilitates accurate recognition and response to human instructions, bridging the gap between AI and humans. Despite the leaps in Natural Language Processing (NLP), a comprehensive grasp of Natural Language Understanding (NLU) remains a formidable challenge.

The growing relevance of NLU is mirrored in its fundamental applications - the development of chatbots and voice-enabled bots. These autonomous entities are revolutionizing our interaction with digital platforms across diverse sectors. Their ability to seamlessly interact and respond to their environment is pivotal to their success and hinges on their proficiency in understanding and interpreting human language.

Intent classification is particularly instrumental in areas such as customer service and sales, enhancing the customer experience by providing timely responses, handling large volumes of queries, and offering personalized services.

This project hones in on intent classification for Slavic languages, emphasizing the Polish language. We draw inspiration from the recently introduced MASSIVE dataset [6], which presents an intriguing challenge of modeling intents across an array of 51 different languages. We were able to fine-tune and test all models on Polish utterances as well as combined utterances of 3 Slavic languages present in the dataset.

Our objective is to create a model proficient in intent classification, capable of assimilating and adapting to the unique features of Slavic languages. The choice to focus on Slavic languages is motivated by their shared characteristics, including morphology, word order, phonetic system, lexicon, verb aspect, complex sentence structures, and accentuation. By leveraging these linguistic parallels in NLU, we aim to develop a model that outperforms those trained on multi-language data.

In addition to adopting multilingual models like XLM [4] and mT5 [26], we also plan to explore the potential of HerBERT [17], a transformer-based model pre-trained specifically on Polish language data. The incorporation of HerBERT is expected to enhance our model's understanding of Polish, thereby improving its overall performance.

Our approach involves evaluating intent accuracy for intent labels. We will compare our findings with the results from the MASSIVE dataset paper, using the XLM, mT5 and HerBERT models as our benchmarks. Through this study, we aim to contribute fresh insights and surpass existing base-

lines in the field of intent classification for Slavic languages.

2 Related work

Over the years, several models have been constructed in pursuit of achieving the best performance in the intent classification field. The first notable one is the BERT model [5]. Even though it was not initially designed for intent classification, it has found its application in this task [2]. It has been later refined through dataset and attention mechanism improvements [12] or by incorporation of additional modules [23].

The recent onset of Large Language Models (LLMs) such as GPT-3.5 or GPT-4 brings forth new tremendous possibilities in various domains [14], also in intent classification. However, while GPT-3.5 demonstrates impressive performance in certain tasks, it still exhibits limitations in others [3], indicating the need for further testing and research to fully adapt LLM capabilities to NLU tasks.

Throughout the years, there has been a noticeable transition towards the utilization of huge, multilingual corpora, which includes datasets like SLURP [1], NLU Evaluation Data [13] and Cross-lingual Multilingual Task Oriented Dialog [22]. Significantly, Amazon's MASSIVE dataset stands out as the most recent addition to this trend, featuring 1 million realistic, labeled virtual assistant utterances spanning 51 languages. Numerous studies show that using several languages leads to significantly better results [9], [10], [15], [16], [25].

The availability of such datasets may prove beneficial for other languages like Polish, as they provide opportunities for enhancing existing baselines such as HerBERT or comparable models utilized in other Slavic languages like Russian [8], Slovak [18] or Czech [24].

Nevertheless, a challenge persists for less-utilized Slavic languages and resource-scarce languages in general. One potentially promising method involves data generation through LLMs [21]. Another interesting solution may lie in modular models, as this approach enables the transfer of learned skills from resource-rich languages to resource-scarce target languages [7], [19].

3 Problem

Our approach involved fine-tuning 5 models: HerBERT, XLM-RoBERTa, XLM-V [11], T5 [20] and mT5. Pre-trained versions of the models were used, augmented with classifications specific to their architecture and then fine-tuned on MASSIVE dataset.

MASSIVE dataset consists of $\sim$16,500 utterances labeled with 60 intents and translated into 51 languages (3 of which were Slavic languages). Models were fine-tuned and tested using two sets of data: Polish utterances and combined utterances from Polish, Slovak and Russian.

4 Experimental Set-up

Implementation of data assembly and fine-tuning is available at https://github.com/Tsilkow/slavic-intent-classification.

4.1 HerBERT

We used the HerBERT model, based on the *allegro/herbert-base-cased*.

Training lasted 10 epochs for both the Polish language dataset and for the combined Slavic languages dataset. The Slavic languages dataset was 3 times larger than the Polish dataset.

We ran multiple experiments on both datasets, manipulating learning rate to get the best result. In the end, we decided on a learning rate of $5e{-}6$ for the Polish language dataset, and $5e{-}5$ to work with the combined Slavic languages dataset. All computations were performed on the Nvidia A100 graphics card.

The HerBERT models are sequence classification models, where input sentences are mapped to a fixed set of classes. The models make use of a classification head at the top of the pre-trained transformer, providing a single-layer perceptron which is fine-tuned from the pre-trained weights.

4.2 T5/mT5

We decided to use the standard T5 model and two mT5 models - mT5-base and mT5-small. mT5 is an enhanced version of T5, pretrained on multilingual dataset. mT5 model is also significantly bigger. mT5-small has nearly 50% more parameters than T5 and mT5-base nearly 3 times more than T5.

The usage of a smaller mT5 model is justified by the fact that all models were fine-tuned for 10 epochs. We experimented with many different learning rates but eventually, 3-e4 was selected for T5, 1e-3 for mT5-small and 5e-4 for mT5-base. We used a batch size of 50 for T5 and 25 for mT5s.

All T5-based models are generative which means they create sentences token by token. Therefore, instead of using a classification head, the labels are generated directly by the model. Loss computation is also a little bit different as it is done by calculating cross entropy between labels and logits outputted by the language modeling head.

4.3 XLM-RoBERTa/XLM-V

We used XLM-RoBERTa and XLM-V, both of size base.

XLM-RoBERTa combines the cross-lingual abilities of XLM with the optimizations of RoBERTa. It's trained on a large amount of multilingual data, and as a result, it has shown strong performance on a wide range of natural language understanding tasks across many different languages.

XLM-V enhances XLM-RoBERTa by utilizing a larger, more language-specific vocabulary and a unique tokenization approach for better semantic meaning, thereby improving performance across low-resourced languages. XLM-V is almost twice as large as XLM-RoBERTa, and therefore, we encountered some GPU RAM issues during the training.

Both models were trained for 10 epochs. Experiments with different learning rate values have shown that both models perform best with a learning rate of $5e{-}5$ on all datasets.

5 Results and Discussion

5.1 HerBERT

HerBERT, being pre-trained specifically on a large Polish corpus, unsurprisingly achieved high performance on the Polish language dataset, with an accuracy of 87.39% after 10 epochs of fine-tuning.

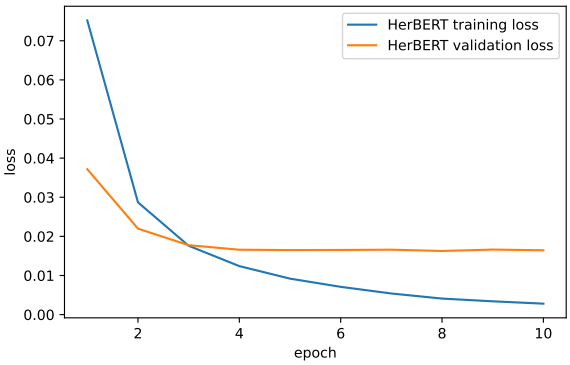

Figure 1: Training and validation loss of HerBERT on the Polish language dataset.

In the more challenging task involving the combined Slavic languages dataset, HerBERT also demonstrated robust performance with an accuracy of 84.52%. This is quite an impressive achievement considering the complexity added by the multilingual nature of the combined dataset.

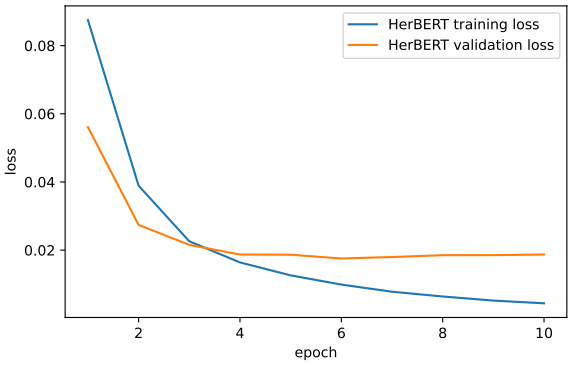

Figure 2: Training and validation loss of HerBERT on the combined Slavic languages dataset.

It's worth noting that HerBERT, given its design and pre-training, naturally performs exceptionally well with Polish language tasks. However, the high accuracy score on the combined Slavic languages dataset underscores its potential as a tool for multilingual tasks within the related Slavic languages, implying a good generalization capability.

These results show that while HerBERT performs excellently in its specialized task (Polish language), it also has potential utility for broader, multilingual tasks, thus demonstrating its flexibility and robustness.

In addition, an important observation we made was the scalability of HerBERT training with respect to the size of the dataset on different types of GPUs. When using an Nvidia A100 graphics card, the learning time for the combined Slavic language dataset was about three times longer than for the Polish language dataset, corresponding proportionally to the difference in their size. This shows almost linear scalability when using high-end computing capabilities such as those provided by the Nvidia A100.

In contrast, when using the Nvidia Tesla T4 graphics card, a popular choice for machine learning workloads but less powerful than the A100, the scalability was less favorable. Specifically, training the combined Slavic languages dataset, which was three times larger, took over eight times longer than training the Polish language dataset, instead of the expected three times.

This non-linear increase in training time with dataset size on the Nvidia Tesla T4 suggests that while the HerBERT model can scale well with powerful hardware like the A100, its performance may be less optimal on less powerful GPUs, particularly for large-scale multilingual tasks.

5.2 T5/mT5

T5 achieved decent results but as expected mT5 was outperforming it due to multilingual pretraining and a bigger number of parameters. Out of the three T5-based models the mT5-small appeared to be the best one.

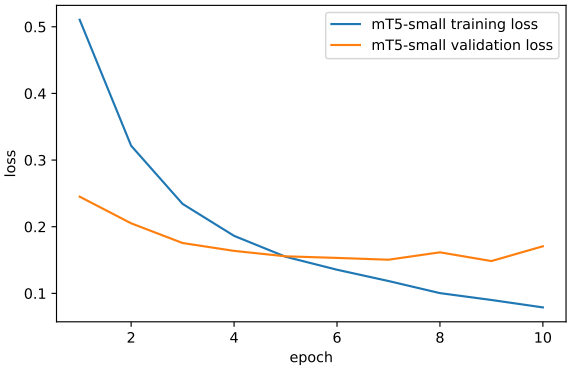

Figure 3: Training and validation loss of mT5 on Polish dataset.

The standard version of mT5 (mT5-base) learned much slower. However, it would likely have emerged as the winner if it had been given more epochs.

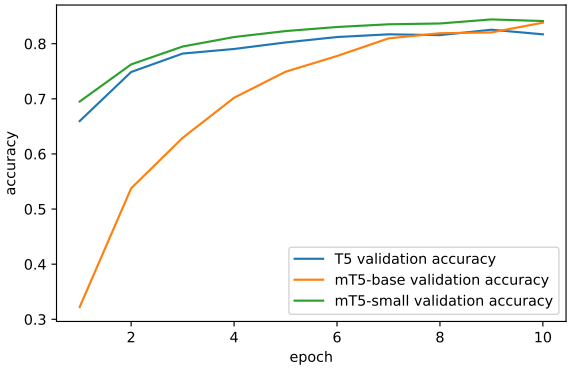

Figure 4: Validation accuracy on Polish dataset of all T5-based models.

Later in the paper when mentioning the results of mT5 we will always refer to the mT5-small as it was slightly better than the mT5-base.

T5	mT5-small	mT5-base
79.39	84.43	83.36

Table 1: Comparison of T5 models accuracy [%] on the test dataset for Polish language after 10 epochs of fine-tuning.

5.3 XLM-RoBERTa/XLM-V

For the Polish language, both models achieved very close results, but XLM-V outperformed XLM-RoBERTa on the test dataset for combined languages.

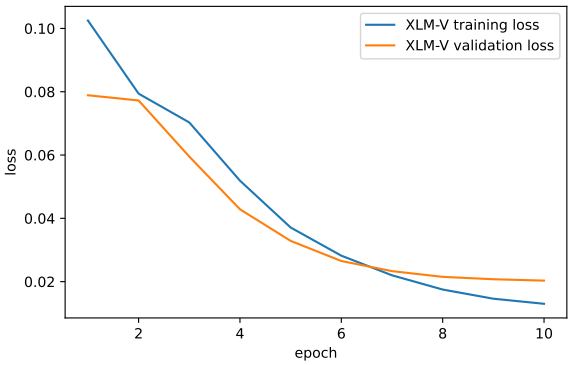

Figure 5: Training and validation loss of XLM-V on Polish dataset.

XLM-V seemed to train a bit slower - this speed could potentially be affected by the differences in the sizes of the models. However, after multiple training runs, we discovered that XLM-V is a bit more robust to changes in learning rate and, therefore, trained more stably than XLM-RoBERTa.

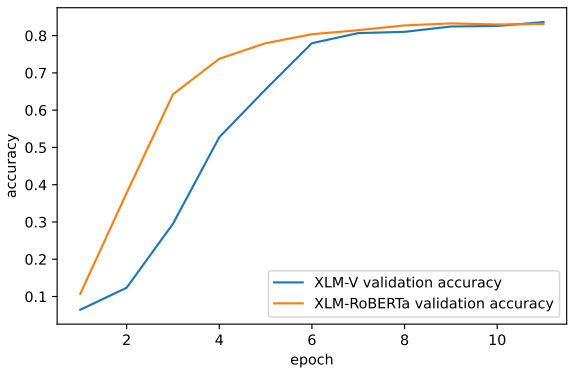

Figure 6: Validation accuracy of XLM-V-base and XLM-RoBERTa-base on Polish dataset.

As XLM-V appeared to be more robust and performed much better on the combined dataset, we consider it the best out of the XLM family, and will therefore refer to it later in this paper.

XLM-RoBERTa-base	XLM-V-base
82.98	82.59

Table 2: Comparison of XLM models accuracy [%] on the test dataset for the Polish language after 10 epochs of fine-tuning.

5.4 Comparison

All models have achieved over 80% accuracy on the test set in both Polish intent classification as well as intent classification in all Slavic languages. Both mT5 and XLM-V scored better in Slavic classification, which is expected due to their multilingual capabilities, while HerBERT scored worse as its pre-trained only in Polish language.

The highest accuracy of 87,22% in intent classification for the Polish language was achieved by HerBERT with a significant advantage over other models. XLM-V scored best in intent classification for Slavic languages with the result of 86,21% however, other models were evaluated at similar levels.

Model	Accuracy [%] - Polish	Accuracy [%] - All Slavic
HerBERT	87.22	84.52
mT5	84.43	84.97
XLM-V	82.59	86.21

Table 3: Comparison of models accuracy [%] on the test split when using dataset with only Polish language and when using combined dataset consisting of 3 Slavic languages: Polish, Slovak and Russian.

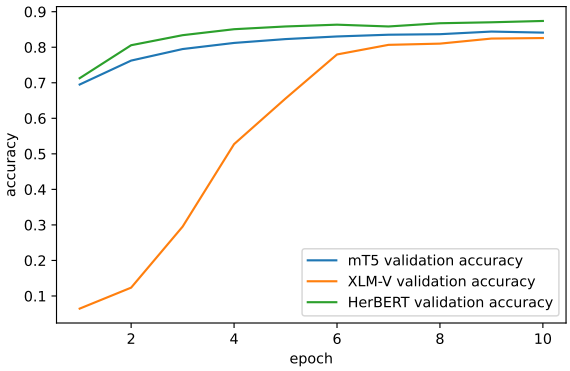

Figure 7: Comparison of accuracy when using dataset with only Polish language.

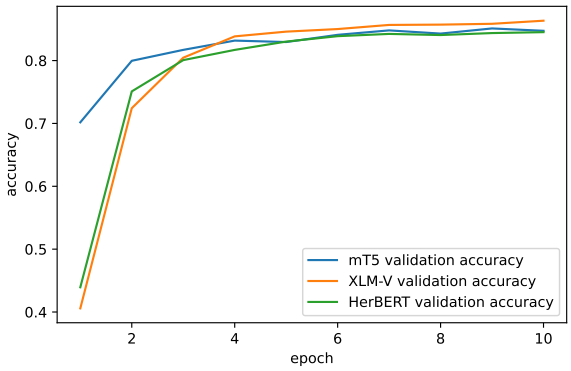

Figure 8: Comparison of accuracy when using dataset consisting of Polish, Slovak and Russian.

6 Conclusions

In this paper, we presented three approaches for intent classification in Polish language and in combined Slavic languages. We fine-tuned and evaluated HerBERT, T5 family and XLM family on MASSIVE dataset and provided results. Training was successful with all models achieving scores of over 80%.

In the T5 family of models, we found that mT5-small was performing the best on our benchmarks. Similarly – between XLM-RoBERTa and XLM-V – the latter one proved superior.

On the task of classification of intent in the Polish language, HerBERT scored the highest, significantly above all other tested models. In a broader task of intent classification in Slavic languages, evaluation values were much closer together, with XLM-V providing the best results.

References

[1] Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package.

[2] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling.

[3] Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding.

[6] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

[7] Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems.

[8] Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

[9] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

[10] Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020. Pre-training via paraphrasing.

[11] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

[12] Han Liu, Siyang Zhao, Xiaotong Zhang, Feng Zhang, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022. A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism.

[13] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents.

[14] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models.

[15] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

[17] Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. Herbert: Efficiently pretrained transformer-based language model for polish.

[18] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. Slovakbert: Slovak masked language model.

[19] Edoardo M. Ponti, Alessandro Sordoni, Yoshua Bengio, and Siva Reddy. 2022. Combining modular skills in multitask learning.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

[21] Gaurav Sahu, Pau Rodriguez, Issam H. Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models.

[22] Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog.

[23] Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. 2021. Enhancing the generalization for intent classification and out-of-domain detection in slu.

[24] Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – czech bert-like model for language representation.

[25] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models.

[26] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

A Training T5 with Polish labels

Usually, when training a model for classification, we do not care about the names of the labels as the most standard approach involves using a classification head assigning numerical values to the classes. However, in the case of T5, the model has to generate the correct sequence of tokens to assign a label (labels may be even more than 10 tokens long). Because we trained T5 on the Polish dataset we intuitively expected that it may be easier for the model to generate the labels in Polish than in English. Therefore, we translated labels into Polish and checked the model's behavior. However, it seems that the translation of the labels did not help and possibly even worsened the training speed, as the Polish labels consisted of more tokens than the ones in English.

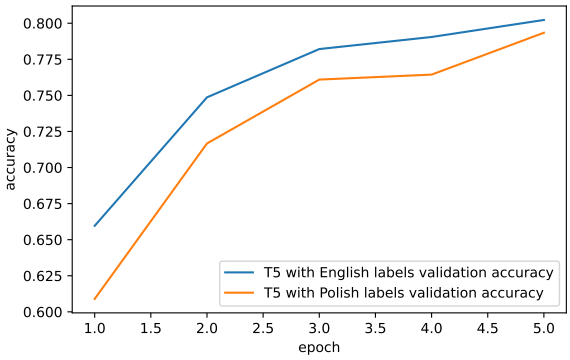

Figure 9: Comparison of validation accuracy of models trained with labels in Polish and in English.

B mT5 results on Slavic languages

Detailed results for mT5-small model on combined dataset consisting of Polish, Slovak and Russian.

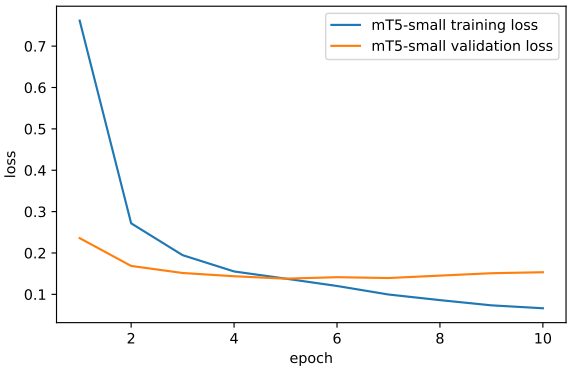

Figure 10: Training and validation loss of mT5 on dataset consisting of Polish, Slovak and Russian.

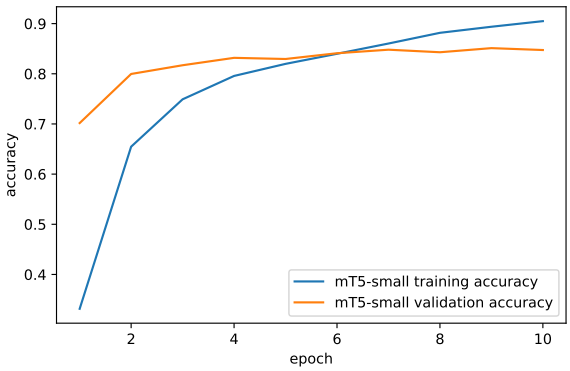

Figure 11: Training and validation accuracy of mT5 on dataset consisting of Polish, Slovak and Russian.

C XLM results on Slavic languages

Detailed results for XLM-RoBERTa and XLM-V models on a combined dataset consisting of Polish, Slovak and Russian.

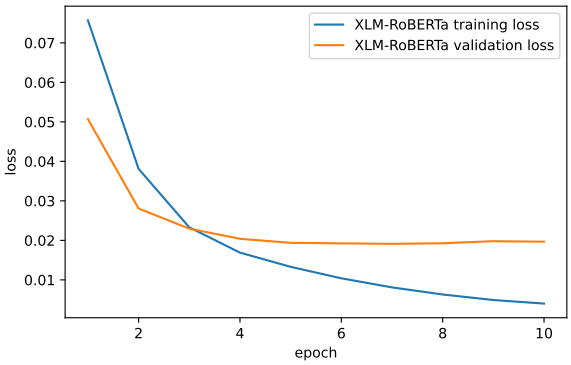

Figure 12: Training and validation loss of XLM-RoBERTa on dataset consisting of Polish, Slovak and Russian.

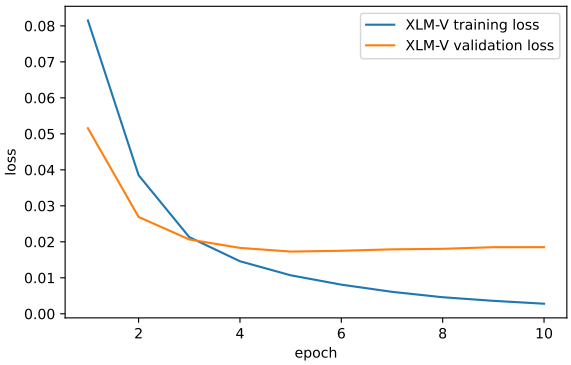

Figure 13: Training and validation loss of XLM-V on dataset consisting of Polish, Slovak and Russian.

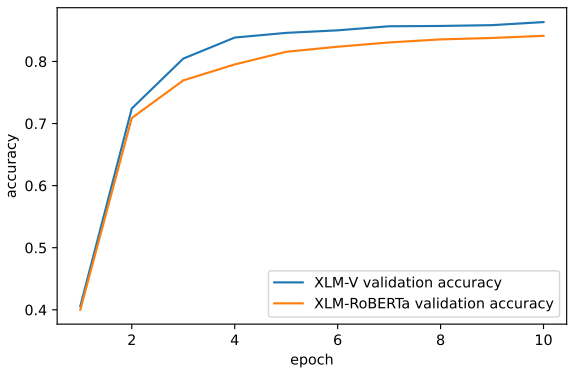

Figure 14: Comparison of validation accuracy of XLM-V and XLM-RoBERTa on dataset consisting of Polish, Slovak and Russian.