Too Many Cooks Spoil the Model: Are Bilingual Models for Slovene Better than a Large Multilingual Model?

Pranaydeep Singh, Aaron Maladry and **Els Lefever**
Ghent University / Groot-Brittanniëlaan 45, 9000 Gent
{pranaydeep.singh, aaron.maladry, els.lefever}@ugent.be

Abstract

This paper investigates whether adding data of typologically closer languages improves the performance of transformer-based models for three different downstream tasks, namely Part-of-Speech tagging, Named Entity Recognition, and Sentiment Analysis, compared to a monolingual and plain multilingual language model. For the presented pilot study, we performed experiments for the use case of Slovene, a low(er)-resourced language belonging to the Slavic language group. The experiments were carried out in a controlled setting, where a monolingual model for Slovene was compared to combined language models containing Slovene, trained with the same amount of Slovene data. The experimental results show that adding typologically closer languages indeed improves the performance of the Slovene language model, and even succeeds in outperforming the large multilingual XLM-RoBERTa model for NER and PoS-tagging. We also reveal that, contrary to intuition, distant or unrelated languages also combine admirably with Slovene, often outperforming XLM-R as well. All the bilingual models used in the experiments are publicly available.[1]

1 Introduction

The last decade has witnessed the increasing popularity of large language models, such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). These transformer-based models have not only pushed the state-of-the-art for a wide range of NLP tasks, but have also shown to perform well in a multilingual setting. Despite their success, these models are also confronted with a number of challenges. First, questions arise regarding their sustainability, given the exponential rise in parameters, and deployability in practical applications. Second, although these models have been shown to achieve good performances for multilingual setups, research has shown that the performance of low(er)-resourced languages, when considering the amount of available Wikipedia data, is below baseline (Wu and Dredze, 2020).

In this research, we want to investigate (1) whether a low(er)-resourced language benefits more from adding data from a typologically closer language, than from more distant languages, and (2) how the performance of such a small dedicated "close family" language model relates to the performance obtained with a purely monolingual model, trained with the same amount of data, on the one hand, and a plain multilingual XLM-RoBERTa model trained with 105 different languages and a huge data set, on the other hand.

For these pilot experiments, we opted to train various transformer-based language models for Slovene, a low(er)-resourced Slavic language. The motivation to perform these "family language model" experiments for a Slavic language originates from the fact that the Slavic languages show a high structural similarity, with a similar inflectional system, and also share a common core vocabulary. As a result, we hypothesize that adding other Slavic languages will boost the performance of the Slovene language model. For each model, we evaluate the performance on three different NLP tasks, namely Part-of-Speech tagging, Named Entity Recognition and Sentiment Analysis.

2 Related research

Deep contextualised multilingual language models, such as mBERT and XLM-R, have shown to perform well for many NLP tasks and for a variety of languages, including low(er)-resourced languages. Nevertheless, previous research has revealed that more similar languages are more helpful for boosting the performance for low(er)-

[1] https://github.com/pranaydeeps/BLAIR

resourced languages. Pires et al. (2019) have investigated the degree to which the representations in Multilingual BERT (Devlin et al., 2019) generalise across languages, by fine-tuning the multilingual model on task-specific data from one language, and evaluating it on another language. Although the authors show that mBERT is able to perform cross-lingual generalization very well, the transfer works best for typologically similar languages, even suggesting that the model works best for languages with similar word orders (Pires et al., 2019). De Vries et al. (2022) performed an extensive transfer learning evaluation with 65 different source languages and 105 target languages, and have shown that, amongst other factors, matching language families, writing systems, word order systems, and lexical-phonetic distance significantly impact the cross-lingual performance.

Multilingual models, such as mBERT and XLM-RoBERTa, use a wide variety of languages from different genera, including the Slavic languages, as part of the same multilingual model. In contrast, other multilingual models have been trained on a smaller selection of languages, with a stronger focus on Slavic languages. The researchers of the DeepPavlov initiative, for example, developed a model for Bulgarian, Czech, Polish and Russian (Arkhipov et al., 2019). While this model was initialised from the multilingual BERT model and then fine-tuned on the task-specific data in the different languages, the CroSloEngual model was pre-trained from scratch for Croatian, Slovene and English and fine-tuned for task-specific data for all languages (Ulčar and Robnik-Šikonja, 2020). This model was built with the intention to apply it for multi- and cross-lingual training, making use of existing data sets for the same task in multiple languages. By doing so, the amount of task-specific data significantly increases, resulting in increased performance of the tasks of NER, POS-tagging and Dependency Parsing. Although this shows that multilingual training causes an increase in performance, the main motivation for the multilingual aspect of the model is the data-hungry nature of the transformer architecture. This same motivation has also led to a transformer model that exclusively uses languages of the Slavic genus. The BERTić language model (Ljubešić and Lauc, 2021) was trained from scratch for Bosnian, Croatian, Montenegrin and Serbian. Whereas the CroSloEngual model uses more distant languages, BERTić selected these languages because they are very closely related, are mutually intelligible and because they are considered part of the same Serbo-Croatian macro language (according to the ISO 639-3 Macrolanguage Mappings). As such, BERTić could be considered not a monolingual or multilingual but rather a macrolingual model. While these languages are exceptionally closely related, this setup does invite the following questions: “How important is the similarity of languages in a combined multilingual language model?” and “Is it preferable to include more closely related languages over distant languages when building a multilingual model?”.

To compare language model performance for similar languages, researchers have often used the World Atlas of Language Structures (WALS) to group typologically similar languages (Yu et al., 2021). WALS (Dryer and Haspelmath, 2013) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive sources such as reference grammars. These linguistic features allow for comparison through qualitative features. This means that they can show in what ways languages are similar and in what ways they differ. However, beside counting the number of shared features, this does not allow for a quantitative comparison. One metric that does allow for a quantitative comparison, i.e. measure *how similar* the languages are, is LDND (Levenshtein Distance Normalized Divided)(Wichmann et al., 2010). This metric was also used by de Vries et al. (2022) in the context of cross-lingual training (training on data from other languages for the same task). Their work has shown that “languages with low LDND distances between source and target language (i.e. when two languages share cognates) are indeed associated with high accuracy, whereas high LDND distances (very dissimilar languages) seem less informative”.

3 System Description

In this research, we want to investigate whether adding data from typologically closer languages improves the performance of a RoBERTa-based language model for three downstream tasks, namely Part-of-Speech tagging, Named Entity Recognition and Sentiment Analysis. To this

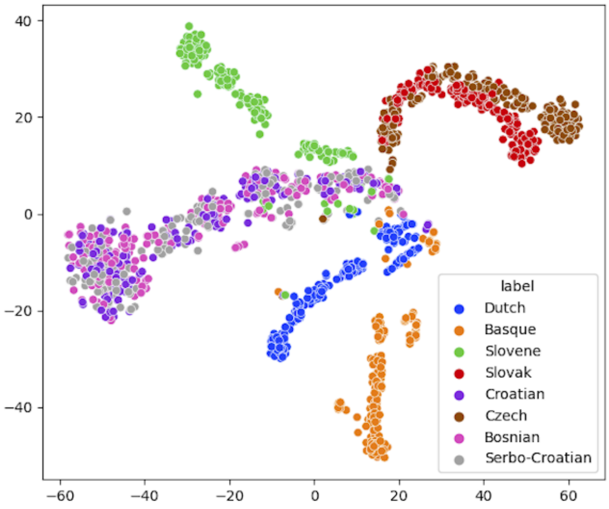

Figure 1: Clustered embeddings from the first layer of XLM-RoBERTa for data from each of the experimental languages visualized with t-SNE. Note that Slovene (in green) is the focal language of this research for which the distance to the other clusters matters most.

end, we performed experiments for the following RoBERTa-based language models including Slovene: (1) a Slovene monolingual model, (2) a Slovene combined with Serbo-Croatian model, (3) a Slovene combined with Slovak model, and (4) a Slovene combined with Czech model. We also performed experiments with two typologically distant languages, Dutch and Basque, for comparison. The motivation for combining specifically these languages with Slovene originates from the LDND measures but can also be linguistically supported. As shown in Table 1, the LDND scores[2] show that Croatian and Serbian are the two closest languages to Slovene. This is in accordance with the fact that these three languages are part of the same sub-group i.e. South-Slavic languages. Croatian is also a neighbouring language. Although Czech (the third-closest language) is not a geographical neighbour and belongs to the West-Slavic sub-group, the areas where Slovene and Czech are spoken share a long (Central European) cultural history (being strongly influenced by developments in the Holy Roman Empire and later the Austro-Hungarian Empire). Therefore, Czech and Slovene can be considered cultural neighbours. Although Slovak has a quite high LDND, the language is mutually intelligible with Czech and shares the same German-dominated cultural history. Therefore, we also included it as one of the languages for our experiments. To evaluate the hypothesis that closely-related languages are more useful for training a multilingual language model than unrelated languages, we also selected two control languages with a high LDND. For this purpose, we found that Basque and Dutch would be good candidates, as their LDND distance is more than twice the distance compared to the Slavic languages[3]. Basque is a completely unrelated language and a prime example of an isolated language that should be sub-optimal for multilingual applications in combination with Slovene. Dutch is part of the same larger Indo-European language family as Slovene, which makes them somewhat, albeit relatively distantly, related. Dutch, therefore, serves as a bridge between related and unrelated languages. These typological distances are also empirically evident in pre-trained multilingual models like mBERT and XLM-RoBERTa. Figure 1 demonstrates the embeddings from the first layer of XLM-R in different languages, visualised using t-SNE (van der Maaten and Hinton, 2008), a dimensionality reduction technique often used for visualising high-dimensional embeddings in 2-dimensions. Similar inferences to the LDND distances can be made using these clusters. Slovak and Czech prove to be quite close. Similarly, Serbo-Croatian, Croatian, and Bosnian also appear to be nearly indistinguishable. Both

[2] Calculated and presented by de Vries et al. (2022).

[3] We only selected languages with Latin script because the difference in the script could potentially increase the difficulty of modeling two languages simultaneously. For our experiments, Serbian data in Latin script was considered as ‘Serbo-Croatian’, meaning that this data does not include Serbian data written in the Cyrillic script.

Language	Distance
Croatian	28.36
Serbian	34.19
Czech	35.68
Bulgarian	40.24
Slovak	44.25
Polish	46.38
Russian	51.63
Ukrainian	52.49
Belarusian	53.85
Basque	100.12
Dutch	90.84

Table 1: LDND distance between Slovene and closely related (Slavic) languages as well as two more distant languages sharing the same Latin script (Basque and Dutch).

	Wiki Data	OSCAR Data
Slovene	276 MB	1 GB
Slovak	300 MB	6 GB
Czech	1 GB	33 GB
Bosnian*	143 MB	165 KB
Croatian*	302 MB	169 MB
Serbo-Croatian*	435 MB	9 MB
Dutch	1.7 GB	47 GB
Basque*	279 MB	503 MB

Table 2: Data sizes of the monolingual corpora used for pre-training the monolingual Slovene baseline and bilingual models. Languages marked with an * have smaller data sizes than Slovene.

the Serbian-Croatian-Bosnian cluster, as well as the Czech-Slovak cluster, are quite close to the Slovene cluster. Dutch and Basque are distantly clustered, with Basque being the farthest of all the visualised languages.

3.1 Experimental setup

As explained, we train bilingual models for Slovene with closely related Slavic languages (Serbo-Croatian, Czech, and Slovak) and with more distant and unrelated languages (Basque and Dutch). To construct monolingual data sources for each of these languages, we use OSCAR 2.0[4] and the latest Wikipedia data dumps[5]. An overview of these sources is summarised in Table 2. Slovene, having a total of 1.276 GB of data serves as the focal point of all the experiments, and therefore data for all the other languages was restricted to the same amount. This allows us to focus the evaluation on the effect of each added language individually and removes data size as a potential variable impacting the performance.

Because of the limited available data and the low LDND distance between Croatian and Serbian (only 19.4), the fact that they are mutually intelligible and considered to be part of the same macro language, we combine the data for Serbian, Croatian and Bosnian to a total data size of 1.06 GB to train a macro-lingual model like BERTić. The data for Basque was also slightly lesser with a combined data size of 782 MB, which might account for some slight disparities. By running the experiments in a controlled setting, viz. evaluating language models built with a very limited data set of similar size, we ensure that the data size is not a variable when drawing inferences from the experiments.

To construct each bilingual model, we combine the data for Slovene (1.276 GB) with the same amount (1.276 GB in size) of randomly selected monolingual data from a second test language, except for Basque (782 MB) and Serbo-Croatian (1.06 GB). After shuffling the combined data, we construct a BPE Tokenizer with 64,000 sub-words and train for the Masked Language Modelling (MLM) objective, using a standard RoBERTa-base architecture, with a max sequence length of 512, starting learning rate of $6e-4$, 3000 warm-up steps and a weight decay of 0.01. We use 32 Nvidia A100 (40 GB) GPUs, with a batch size of 32 per device, and gradient accumulation for 8 steps, thus adding up to an effective batch size of 8192. The AdamW optimizer was used for optimisation with an epsilon of $1e-6$, a β_1 value of 0.9, and a β_2 value of 0.98. All the bilingual models were trained for 30 epochs, or approximately 60,000 steps, which took approximately 40 hrs per model.

Finally, we also train a monolingual Slovene model with only the base 1.276 GB of Slovene data, with identical hyper-parameters, except restricting the vocabulary to 32,000 to account for only having a single language. The monolingual model is intended to serve as a benchmark to quantify the potential improvements obtained by adding the secondary test language in combination with Slovene.

[4] https://huggingface.co/datasets/oscar-corpus/OSCAR-2109

[5] https://dumps.wikimedia.org/backup-index.html

4 Evaluation and Discussion

We evaluated the various versions of the RoBERTa language model on three different downstream tasks: one semantic task, being Sentiment Analysis, one syntactic task, being Part-of-Speech (POS) Tagging, and one task requiring both syntactic and semantic understanding, namely Named Entity Recognition (NER). For Sentiment Analysis, we use the SentiNews dataset (Bučar et al., 2018), which consists of news documents annotated with three sentiment labels (neutral, positive, and negative). We use the sentence-level sentiment setup with approximately 169,000 sentences, distributed into 80:10:10 for training, validation, and testing, respectively. For NER we, use the WikiANN (Rahimi et al., 2019) dataset with 15,000 train samples, and 10,000 samples each for validation and testing. Finally, for POS Tagging, the SSJ Treebank part of the Universal Dependencies[6] project is used, consisting of 13,000 annotated sentences, split into an 80:10:10 setup for training, validation and testing as well. For all downstream tasks, the respective RoBERTa models were fine-tuned for 10 epochs, with a learning rate of $5e-5$ with 500 warmup steps followed by a linear weight decay of 0.01. The results are summarised in Table 3.

Firstly, the Monolingual Slovene model seems to perform comparably to XLM-Roberta on all tasks, while only performing slightly worse than the Upper-Bound (UB) SloBERTa[7] model, which was trained on significantly (21 times) more data. This indicates that the presence of the additional 99 languages does not have a significant impact on Slovene performance. The bilingual model with Slovene+Serbo-Croatian seems to perform the best for NER, even outperforming the state-of-the-art SloBERTa (UB), while the Slovene+Czech model seems to be the best for POS Tagging, and only 0.06% worse than the UB, while the Slovene-Slovak model works best for Sentiment Analysis. For all three tasks, the best models come from the typologically closely related languages, however, the models with distant languages, Slovene+Dutch and Slovene+Basque, do not perform as badly as hypothesized. Both models outperform the monolingual baseline, while sometimes also competing with the closely related languages in some settings. This is an interesting and rather counter-intuitive finding, since it suggests that the addition of data, irrespective of the language, is helpful for a given target language. Even Basque, with an LDND distance of more than 100, is able to influence the Slovene performance in a positive sense. This incites the following question: If all languages are indeed useful, irrespective of their differences, why is XLM-RoBERTa the worst performing model then, with the highest amount of combined data? A logical inference would then be that after a certain amount of languages, the representation power of the RoBERTa-base setup is not sufficient to model all 104 languages simultaneously, resulting in degradation for the poorly represented languages in the data, as would be the case for Slovene. These observations might still indicate, however, that a multilingual model with 3 or more languages might show further improvements to our bilingual setup.

In general, one can observe a downward trend as we move further away from Slovene in terms of typological similarity or LDND distance. This trend can be seen more clearly in Figure 2 for POS and NER, while the trend is not as explicit for Sentiment Analysis, with a few anomalies. We dive further into the potential reasons for the inconsistencies with Sentiment Analysis performance in the next section.

5 Manual evaluation for Sentiment Analysis

As the results for the sentiment analysis task do not align with our hypothesis and do not follow the tendencies we noticed for the other tasks, we decided to have a look at the predicted labels to find an explanation for these deviant results.

A closer look at the evaluation data revealed a couple of reasons for the unexpected results. The evaluation data was selected from curated economic and political news corpora, characterised by a more neutral writing style. As a result, the sentiment is often implicit or ambiguous and requires world knowledge and human experience to be interpreted correctly. This is also confirmed by the modest inter-annotator agreement reported by Bučar et al. (2018), with F1-scores below 65% for their 3-way classification models.

As shown in Example 1, a rather neutral statement can also carry an implicit (negative) sentiment although it was annotated as neutral in the data set.

[6] https://universaldependencies.org/

[7] https://huggingface.co/EMBEDDIA/sloberta

	UB	SL-SBC	SL-CS	SL+SK	SL+NL	SL+EU	Monolingual	XLM
NER	0.9410	**0.9441**	0.9422	0.9425	0.9396	0.9406	0.9396	0.9409
POS	0.9902	0.9892	**0.9896**	0.9887	0.9892	0.9889	0.9878	0.9865
Sentiment	0.6835	0.6633	0.6660	**0.6757**	0.6657	0.6628	0.5925	0.6664

Table 3: F1-scores for the tasks of NER, POS-tagging, and Sentiment Analysis. The Upper-Bound (UB) is the monolingual SloBERTa model, trained with 21 times more monolingual data compared to our monolingual Slovene RoBERTa baseline model).

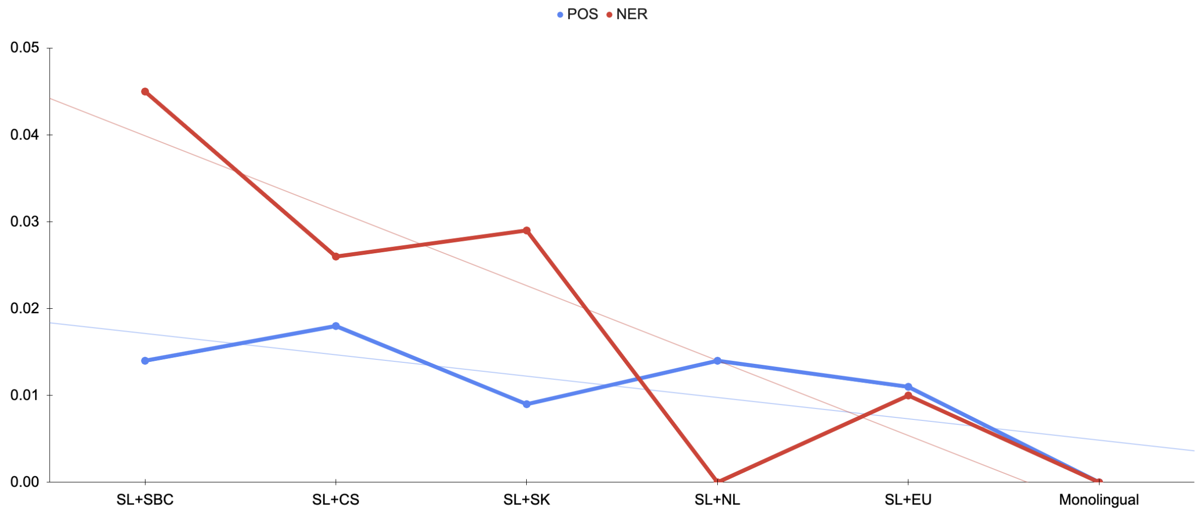

Figure 2: Differences in F1-score of all evaluated models compared to the monolingual baseline. The models are listed on the X-axis in ascending order of linguistic distance of the second language (in relation to Slovene). The monolingual baseline is included for completeness.

Example 1

Več kot milijon Parižanov se je ***moralo*** *v službo odpraviti kar peš ali s kolesom*

(translation: More than a million Parisians ***had to go*** *to work on foot or by bicycle)*

A second cause for errors is that the annotators also took the context into account for labeling the sentiment of individual sentences. This can cause a contextual sentiment to seep into the label of a rather neutral sentence. In Example 2, a neutral sentence was tagged as "positive", although this cannot be inferred from the sentence itself.

Example 2

SI: *Vsak bo tako prispeval polovico zneska.*

(translation: Each will thus contribute half of the amount.)

In some cases, the erroneous sentence splitting of news articles resulted in single-word sentences (named entities and numbers), as shown by the following examples:

Example 3

Lukea Koper

Intereuropa

Gorenje

KRKA

1,75%

While these single-word sentences should be neutral, they were still annotated with a positive or negative sentiment (most likely due to the context again).

In order to get a general idea of how the bilingual models compare to the monolingual Slovene model, we performed a shallow evaluation of the results. Considering the complexity of the task, we focused on samples that were not predicted or annotated as neutral. This way, we get an indication of the performance on more explicit sentiments. This evaluation has underlined the improvement of both the Dutch+Slovene and Serbo-Croatian+Slovene models over the monolingual model (which, in turn, generally outperforms the multilingual model). In Examples 4, 5, and 6, both bilingual models with Dutch and Serbo-Croatian predict a correct sentiment, whereas the monolingual model fails.

Example 4

Nižji dobiček ameriških podjetij

(translation: Lower profits for US companies)

Example 5

Najbolj je padla prodaja oblačil in tehničnega blaga.

(translation: Sales of clothing and technical goods fell the most)

Example 6

Ko ugotoviš, da si pogumna oseba, lahko premagaš strah in neuspeh.

(translation: When you realize that you are a brave person, you can overcome fear and failure.)

When comparing non-neutral sentences where these two bilingual models disagree, it becomes a lot harder to find tendencies. In some cases where the sentiment is more explicit, the Serbo-Croatian+Slovene model provides a more intuitive prediction, as shown in Example 7, 8, 9. However, more analysis and further statistical evidence is needed to support this hypothesis.

Example 7

Najprej nekaj besed o Jožetu Pučniku: voditelj demokratične opozicije Slovenije je na svoji koži izkusil surovost prejšnjega režima, sedem let je bil v zaporu zaradi "subverzivne dejavnosti".

(translation: First, a few words about Jože Pučnik: the leader of Slovenia's democratic opposition experienced the cruelty of the previous regime firsthand, he was in prison for seven years for "subversive activity".)

Bilingual Dutch prediction: Positive

Bilingual Serbo-Croatian Prediction: Negative

Example 8

Sama sebe sem bodrila, res mi je odleglo.

(translation: I cheered myself up, I was really relieved.)

Bilingual Dutch prediction: Negative

Bilingual Serbo-Croatian Prediction: Positive

Example 9

Tudi Petrol je cenejši za skoraj tri odstotke.

(translation: Petrol is also cheaper by almost three percent.)

Bilingual Dutch prediction: Negative

Bilingual Serbo-Croatian Prediction: Positive

6 Conclusion

This paper presents a pilot study to investigate whether adding data from typologically close languages improves the performance of a monolingual model for a low-resourced language, Slovene in this case. To summarise the results, our experiments showed that adding data from a second language always helps, even if this language is more distant. In addition, the trained bilingual models outperform the very large multilingual model in almost all cases. Finally, the bilingual Slavic models outperform the bilingual models with more distant languages for the task of Named Entity Recognition and POS Tagging barring a few anomalies, whereas this is not confirmed for the task of Sentiment Analysis. As the results for Sentiment Analysis were somewhat counter-intuitive and not in line with the findings of the other tasks, we decided to also perform a small manual analysis where we outlined a number of issues with the complexity and subjectivity of the sentiment analysis task, including modest inter-annotator agreement and a number of ambiguous instances.

In future research, we will perform validation experiments for additional combinations and downstream tasks, especially because the deviant scores for Sentiment Analysis might be partly due to the nature of the evaluation set used. Additionally, it would also be worthwhile to check whether adding additional data for a second language (like Croatian) would have a stronger positive impact on the evaluation of Slovene compared to adding the same amount of data for a third language (Czech). Finally, we will also investigate simultaneously adding more than two languages to the training setup, to find the optimal inflection point for multilingual setups, after which some performance degradation is likely.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52:895–919.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7676–7685. Association for Computational Linguistics.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Nikola Ljubešić and Davor Lauc. 2021. Berti\'c–the transformer language model for bosnian, croatian, montenegrin and serbian. *arXiv preprint arXiv:2104.09243*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 104–111. Springer.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Søren Wichmann, Eric W Holman, Dik Bakker, and Cecil H Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, 389(17):3632–3639.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *CoRR*, abs/2005.09093.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7210–7225, Online. Association for Computational Linguistics.

7 Limitations

The primary limitation of this work is that the hypothesis can be validated for more languages, tasks, and typological families. However, it takes a lot of computational resources (1280 GPU hours on Tesla A100 GPUs) and training time, to train and validate each model, thus having quite a large carbon footprint (approximately 85kg of CO_2 emission per model). The results are also not consistent for the task of Sentiment Analysis but this can be accounted for by the issues mentioned in Section 5. The tasks, while being varied (in a semantic and syntactic sense), might not cover general language understanding as well as comprehensive benchmarks like GLUE. However, since we attempt to validate the hypothesis for under-resourced languages, large benchmarks are often hard to come by.