Cross-lingual Sentiment Analysis of Official EU Slavic Languages

Gaurish Thakkar, Nives Mikelić Preradović, Marko Tadić – Faculty of Humanities and Social Sciences – University of Zagreb

Cleopatra Final Event 2023-05-15/16 Hannover, Germany

Introduction

Sentiment Analysis (SA) is the field of study that analyses the opinions, sentiments, appraisals, attitudes, and emotions expressed in the written text regarding entities and their attributes" (B. Liu, 2012). While the availability of annotated resources immediately contributes to higher performance in high-resource languages, data scarcity is a problem in low-resource languages. To produce reasonable text categorization results, current state-of-the-art approaches rely on transformers (Devlin, 2019). However, the sheer number of models available publicly in model hubs is enormous and poses the question of initial language model selection. In this study we probe existing language models and propose methodology for candidate selection that is then trained using multi-task joint training fashion.

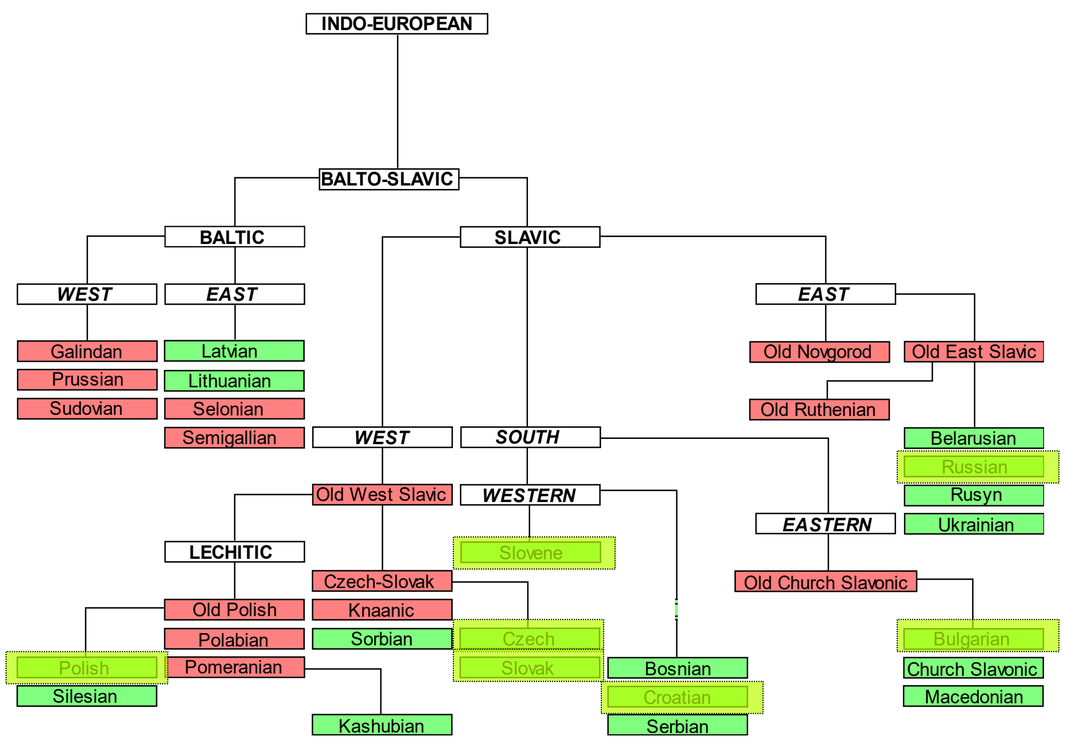


Fig.1. Indo-European family tree.

Aims

- Probe existing pretrained language models for a suitable candidate that will improve low-resource language performance.
- Improve sentiment analysis classification models for low-resource languages using resources from the same as well as distant family languages.

Dataset

Our supervised resources include datasets in eight distinct languages, seven of which are official EU languages. We considered English to be the source language for all pairs of languages. Bulgarian, Croatian, Czech, Polish, Slovak, and Slovene are the target languages.

Language	Dataset	Train	Val	Test
Bulgarian	Cinexio	5520	614	682
Croatian	Pauza	2277		1033
Czech	CSFD	63966	13707	13707
English	MARC	200,000	5000	5000
Polish	all2	28581	3572	3572
Russian	ROIMP 2012	4000	260	5500
Slovak	Reviews3	3834	661	1235
Slovene	KKS	3977	200	600

table.1. Distribution of sentiment analysis datasets.

Methodology

1. Probing

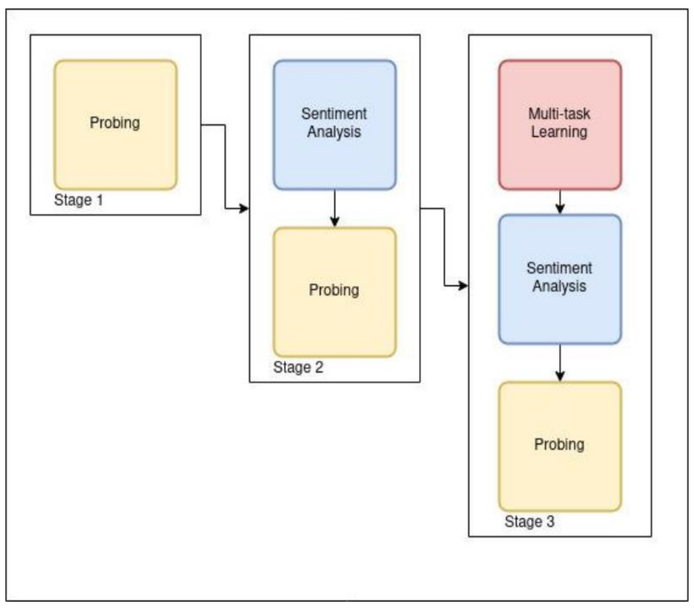
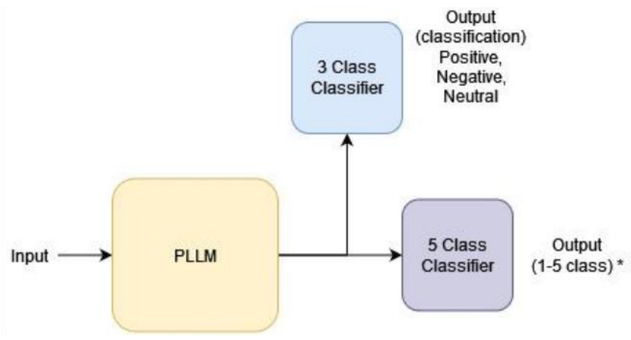


Fig.2. Probing.

We probed the language model using negation, bitext, and paraphrase datasets from the target languages and correlated the performance with sentiment analysis scores.

2. Joint-training



For each language, a dataset from the target language is:

- used directly to train the model (like Bulgarian).
- combined with a single dataset from a distant language family (like English).
- combined with a single dataset from a different subbranch of the same language family (like Russian, Polish, or Czech).
- merged with several low-resource language datasets (Croatian, Slovak, and Slovene)

Results

We find that task of negation bears moderate correlation (Spearman rank correlation coefficient ϱ) ~ 0.38 with the sentiment analysis score. XLM-R-base performs best in the majority of cases. High-resource languages do not benefit from joint training. Low-resource languages show statistical improvement when data from the same family as well as distant families is combined, which ranges from 1%-10% for various languages.

Conclusion

The proposed methodology allows probing and language model selection using the task of negation. In the absence of large dataset, joint-training using the data from the same family as well as distant family can be leveraged to improve the sentiment analysis scores.

References

Liu, B., & Zhang, L. (2012). A survey of opinion mining and s entiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997.