FRI Natural language processing course 2024
Project report

University *of Ljubljana*
Faculty *of Computer and*
Information Science

Unsupervised Domain adaptation for Sentence Classification

Bernard Kuchler, Jan Šuklje, and Luka Šveigl

Abstract

Advisors: Boshko Koloski

Introduction

We seek to improve article representation in specialized domain articles by adapting a pretrained multilingual SBERT model, which is not tuned to specific fields. The non tuned model will serve as a baseline to improve upon. We will fine tune the SBERT model on articles from the SentiNews database[1]. We will attempt to accomplish this by using two adaptation techniques: TSDAE (Transformer-based Denoising AutoEncoder) and GPL (generative pseudo labeling).

SBERT, created by Reimers and Gurevych [2] is an improvement upon BERT for the creation of semantically meaningful sentence embeddings. It will serve as a good baseline as it can encode semantics for sentences making it easier to compare them. SBERT presents similar sentences with embeddings which are close together in the vector space while dissimilar sentences are far apart. We have used multilingual SBERT models so that we did not require additional training of the model for the Slovene language. We have decided to use and compare the Language-agnostic BERT Sentence Embedding (LaBSE) created by Feng et al. [3] and the paraphrase multilingual MiniLM L12 v2 model [4]. For the baseline performance we adapted the mentioned models for article keyword classification by training a feed-forward neural network on their outputs.

TSDAE, created by Wang et al. [5] is a domain adaptation and pre-training method. TSDAE is used to introduce noise to input sequences by deleting or swapping tokens and turning them into sentence vectors. It then attempts to reconstruct the original input sequence from the vectors. This allows the model to better capture the characteristics of the target domain.

GPL, created by Wang et al.[6] is also a unsupervised domain adaptation method. It consists out of query generation, negative mining and pseudo labeling in a data preparation step and a fine-tuning step.

Methods

Data analysis

In an attempt to better understand our dataset, we performed a simple analysis of the train and test splits. Our main focus was the distribution of keywords pertaining to the articles, which would provide insights into the eventual training procedure. In Figure 1 we can observe the 20 most common keywords in both the training and test splits. Immediately, we can note the massive skew towards keywords ”gospodarstvo” and ”ekonomija”, which suggests our model might have problems with overfitting to these specific categories. More accurately, the keyword ”gospodarstvo” represents a 12.35 percent share in the training set and a 13.4 percent share in the test set. The keyword ”ekonomija” represents a 6.8 percent share in both the train and test sets. Together, we can observe these keywords represent approximately one fifth of the entire dataset. Additionally, certain keywords were present only in the training set, while some others were present only in the test set, which proved to be problematic in the label encoding step.

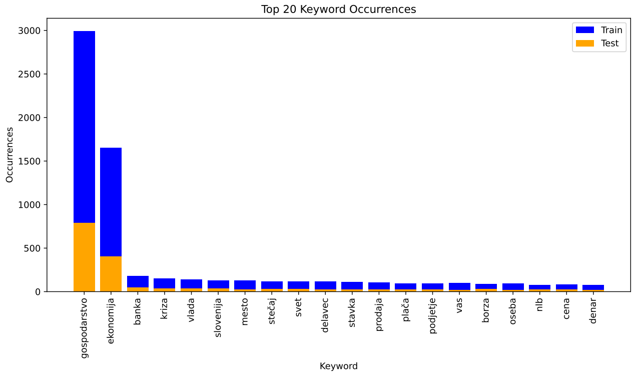

Figure 1. Top 20 most common keywords per dataset.

Baseline model

For the baseline model on which to compare our improvements in article representation in specialized domain articles, we first had to adapt the selected SBERT models for article keyword classification. We achieved this by training a feed-forward neural network on the outputs of the the Language-agnostic BERT Sentence Embedding (LaBSE) created by Feng et al. [3] and the paraphrase multilingual MiniLM L12 v2 model [4]. The training and fine tuning of the feed-forward neural network parameters was performed on the SentiNews database [1] by using the training and testing split. The feed-forward neural network has a multi-label classification output as an article can be given multiple keywords.

We then selected the best performing combination of the trained neural network and SBERT model as our baseline to improve upon in the continuation. The performance of our baseline model can be seen in section Baseline results.

Results

Baseline results

Our first task was determining which sentence transformers perform best on the task of classification of Slovene articles. We mainly focused on testing two sentence transformers: LaBSE and MiniLM-L12-v2. We tested the transformers along with the classification head on multiple different parameter combinations, classification head architectures, learning rates and schedulers, loss functions and optimizers. As we were dealing with a multi-label classification problem, we evaluated the models using the F1, precision and recall scores, alongside Jaccard index. The top 5 results of our evaluations are presented in Table 1, while the testing configurations are presented in Table 2.

Configuration	F1 score	Jaccard index
1	0.450	0.421
2	0.449	0.418
3	0.410	0.358
4	0.409	0.358
5	0.408	0.358

Table 1. Results of the evaluations.

In Table 2 we can observe that the best performing sentence transformer was the LaBSE model, reaching the F1 score of 0.45 and Jaccard index of 0.421. The resulting performance of the baseline model is influenced by the huge skew of the dataset, presented in Figure 1, more specifically, by the keywords ”gospodarstvo” and ”ekonomija”, which occur the most in the articles, as mentioned in Section Data analysis. Due to this the model overfitts on these keywords, resulting in them being predicted more often and in cases where they don’t appear. The results obtained by the evaluations are quite good, especially when taking into account the challenges of multi-label classification, such as much larger complexity of the output space, ambiguity and label dependencies along with data sparsity.

Based on the results, we decided to use the LaBSE sentence transformer for the future parts of this project.

Discussion

Acknowledgments

References

[1] Boshko Koloski, Matej Martinc, Ilija Tavchioski, Blaž Škrlj, and Senja Pollak. Slovenian keyword extraction dataset from SentiNews 1.0, 2022. Slovenian language resource repository CLARIN.SI.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2022.

[4] paraphrase-multilingual-MiniLM-L12-v2 model.

[5] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*, 2021.

[6] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577*, 2021.

Config	Model	Number of layers	Neurons	Loss	Optimizer	λ	Threshold	Scheduler
1	LaBSE	2	1024	BCEWithLogits	ADAM	0.01	0.3	None
2	LaBSE	2	512	multiLabelSoftMargin	ADAM	0.01	0.3	None
3	MiniLM	2	512	multiLabelSoftMargin	ADAM	0.01	0.3	None
4	MiniLM	2	1024	BCEWithLogits	ADAM	0.01	0.3	None
5	LaBSE	2	512	multiLabelSoftMargin	ADAM	0.1	0.533	None

Table 2. Caption