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

Unsupervised Domain adaptation for Sentence Classification

Bernard Kuchler, Jan Šuklje, and Luka Šveigl

Abstract

This report examines unsupervised domain adaptation techniques to enhance sentence classification using a multilingual SBERT model. The goal was to improve keyword and sentiment prediction in specialized domains, utilizing the SentiNews databases. Two methods, Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL), were applied. Two sentence transformers in LaBSE and paraphrase multilingual MiniLM L12 v2 models were compared. Fine-tuning with TSDAE significantly improved keyword classification, and GPL also showed notable enhancements. The LaBSE model performed best as the baseline, highlighting the effectiveness of these adaptation techniques in multilingual sentence classification.

Keywords
GPL, LaBSE, MiniLM, Multi-label model, Multi-task model, TSDAE

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. Our aim is for the model to be able to predict both article keywords and their sentiment.

The non tuned model served as a baseline to improve upon. The SBERT model was then fine tuned on articles from the SentiNews databases created by: Koloski et al. [1] for article keywords and Bučar [2] for sentiment. This was accomplished by using two adaptation techniques: TSDAE (Transformer-based Denoising AutoEncoder) and GPL (generative pseudo labeling).

SBERT, created by Reimers and Gurevych [3] is an improvement upon BERT for the creation of semantically meaningful sentence embeddings. It serves 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 further 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. [4] and the paraphrase multilingual MiniLM L12 v2 model [5]. For the baseline performance, we adapted the mentioned models for article keyword and sentiment classification by training feed-forward neural networks on their outputs.

TSDAE, created by Wang et al. [6] 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.[7] is also an 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

We first analyzed the dataset with annotated article keywords [1]. In an attempt to better understand this 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.

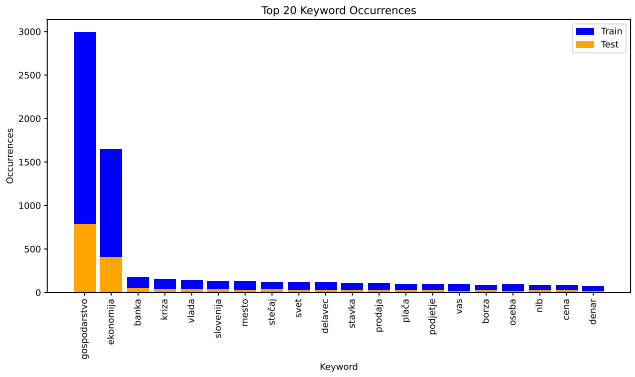

Figure 1. Top 20 most common keywords per split in the SentiNews keywords dataset.

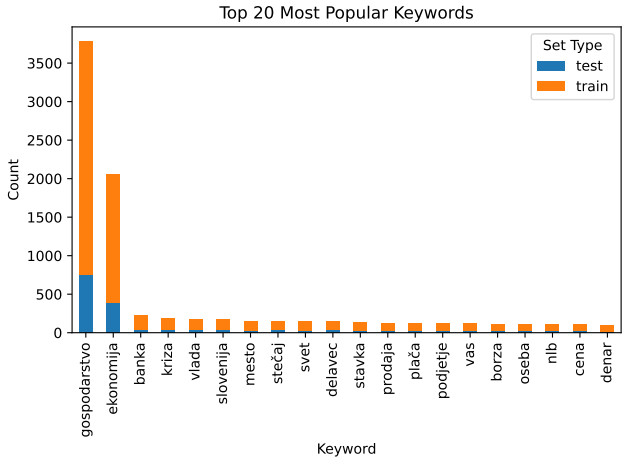

Figure 2. Top 20 most common keywords per split in the SentiNews sentiment dataset.

As we later adapted our model for the task of multi-task classification, we then analyzed the dataset with article sentiment annotation [2]. Here we again performed an analysis of the train and test splits. In Figure 3 we can observe that the majority of the news articles have a neutral sentiment, which would be expected from objective reporting and the use of factual statements. The distribution of the news article sentiments among the two datasets seems to be well balanced, which should have a positive impact on model training and its subsequent evaluation. Additionally, in Figure 2, we can observe that the distribution of keywords in the SentiNews sentiment dataset is similar to the distribution of keywords in the SentiNews keywords dataset.

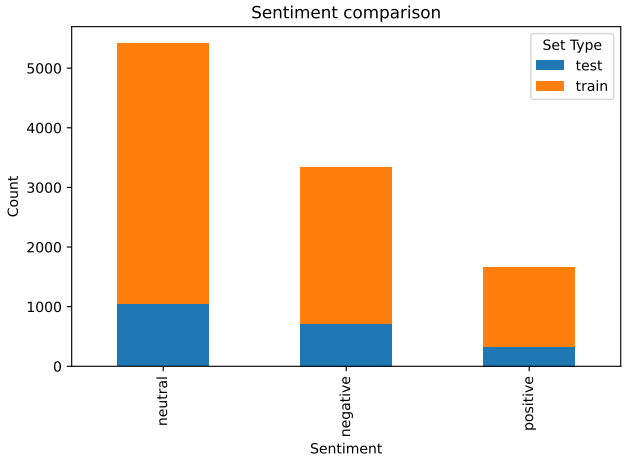

Figure 3. Comparison of sentiment distribution among the articles per split in the SentiNews sentiment dataset.

Baseline model

We first adapted the selected SBERT models for article keyword classification to see what baseline results we could achieve when performing multi-label classification alone. This kind of classification is a harder task when compared to sentiment classification where a single sentiment is predicted per news article. We adapted the selected SBERT models by training a feed-forward neural network on the outputs of the Language-agnostic BERT Sentence Embedding (LaBSE) created by Feng et al. [4] and the paraphrase multilingual MiniLM L12 v2 model [5]. 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 try and improve upon in the continuation. The performance of our baseline model can be seen in the section Baseline results.

Multi-task baseline model

After obtaining the results for article keyword classification, we adapted the best performing model to additionally perform sentiment classification, resulting in a multi-task model. We performed this by adding and training a feed-forward neural network on the outputs of the chosen LaBSE model. The training and fine tuning of the feed-forward neural network parameters was performed on the SentiNews sentiment database [2] by using the training and testing split. This added feed-forward neural network classifier consists of the same classification head for both sentiment and keyword prediction, with different output fully connected layers.

TSDAE

We selected the LaBSE model for further adaptation to improve its performance in our chosen domain, as it performed

best when selecting and training the baseline model. We first adapted the LaBSE model for the news article domain, using the TSDAE adaptation technique. The training data was first chunked in a way that the body of each news article was split into multiple parts of 512 tokens, which is the largest input size to be used by the LaBSE model. Here we then selected 20000 of the mentioned chunks from our training data and added noise to them by deleting tokens with a probability of 60%. We selected the AdamW optimizer to be used during training and fine-tuned its parameters. Once TSDAE has been trained we applied it to the output of the LaBSE model and continued training the combined model. Here we took the fine-tuned parameters selected during the baseline models training to obtain a better comparison of the adaptation technique. We selected the training and testing datasets used for both keyword and sentiment prediction and trained the multi-task model with TSDAE applied to its output.

During the training of TSDAE we selected the following training parameters to be used by the AdamW optimizer. The number of epochs has been set to 5 as larger values had significantly increased training time. We then compared the performance of using different schedulers and selected warmuplinear as it had achieved the best performance for the given parameters. Using weight decay only decreased the performance of the model due to which it was set to 0. The optimizer learning rate was set to 0.0001.

GPL

As with the TSDAE fine-tuning method, we decided to fine-tune the LaBSE model on our SentiNews sentiment dataset, as it performed best when evaluating our baseline model. In our implementation of the GPL fine-tuning method, we split the procedure into 2 separate parts: dataset generation and model fine-tuning. The reason for this was that triplet generation needs to only be performed once, which speeds up the training process of the model, allowing us to modify the training parameters without regenerating the dataset.

To implement the GPl method, we followed the instructions available on the Pinecone website, implementing the Query generation, Negative mining and the Pseudo-labeling steps.

In the query generation step, the method first generates queries from passages present in our text. For this task, we split our article bodies into sentences and passed them into the bkoloski/slv_doc2query model, which generated the necessary passage queries. For the negative mining and pseudo-labeling steps, we selected our target LaBSE model.

Once our training data was generated, we fine-tuned the LaBSE sentence transformer using various combinations of parameters, such as size of training data, number of epochs, learning rate, etc. which will be presented in the Section Fine-tuned 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 the Jaccard index. The top 5 results of our evaluations are presented in Table 1, while the testing configurations are presented in Table 4.

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 of the multi-label classifier.

In Table 4 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 the 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.

Once we adapted our classifier for multi-task classification, we evaluated it using the best performing combination from our baseline model search. This was done to obtain starting results, with which we would compare the results of our fine-tuned implementations. The best results of this evaluation were an F1 score of 0.08133589, Jaccard score of 0.07861936 and sentiment classification accuracy of 0.65388302. We can observe that switching to a multi-task classification problem and a different choice of dataset lowered our keyword classification scores compared to the multi-label classifier.

Fine-tuned results

In Table 2, we present the results in terms of F1 score and Jaccard index for keyword classification and Accuracy for sentiment classification. The TSDAE fine-tuning configurations for the LaBSE sentence transformer are presented in Table 5. We can observe that using the TSDAE fine-tuning method

on the LaBSE sentence transformer more than doubled the F1 scores and Jaccard scores, meaning our model performed much better on the task of classifying keywords. In contrast, the classification accuracy of sentiment decreased slightly with the fine-tuned sentence transformer.

Configuration	F1 score	Jaccard index	Accuracy
1	0.201	0.189	0.631
2	0.192	0.174	0.649
3	0.168	0.146	0.651
4	0.223	0.207	0.648
5	0.233	0.221	0.637

Table 2. Results of the evaluations of the multi-task classifier using the TSDAE adaptation technique.

From the results, it is evident that Configuration 5 achieved the highest F1 score and Jaccard index, indicating the best performance in keyword classification among the TSDAE fine-tuning configurations. However, this configuration also had one of the lowest sentiment classification accuracies, highlighting a trade-off between keyword and sentiment classification performance. This suggests that while TSDAE significantly enhances keyword classification, fine-tuning for optimal performance across multiple tasks remains a challenge.

In Table 3, we present the results in terms of F1 score and Jaccard index for keyword classification and Accuracy for sentiment classification. The GPL fine-tuning configurations for the LaBSE sentence transformer are presented in Table 6. We can observe that the GPL fine-tuning method of the LaBSE sentence transformer yielded slightly better results for the task of keyword classification, once again at the expense of classification accuracy of sentiment.

Configuration	F1 score	Jaccard index	Accuracy
1	0.104	0.104	0.628
2	0.098	0.096	0.640
3	0.288	0.288	0.504
4	0.110	0.108	0.641
5	0.118	0.115	0.631

Table 3. Results of the evaluations of the multi-task classifier using the GPL adaptation technique.

The GPL adaptation technique demonstrated its potential with Configuration 3, which achieved the highest F1 score and Jaccard index among the configurations tested. This indicates a notable improvement in keyword classification over the baseline, though it came with a significant drop in sentiment classification accuracy. The performance discrepancy suggests that while GPL can improve keyword classification, it might require more sophisticated balancing techniques to maintain sentiment classification accuracy.

Discussion

This study demonstrates that unsupervised domain adaptation techniques can significantly enhance sentence classification in specialized domains. By fine-tuning a multilingual SBERT model with Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL), we achieved notable improvements in keyword and sentiment prediction.

TSDAE was particularly effective, markedly improving keyword classification by better capturing domain-specific features. While GPL also improved performance, its impact was less pronounced, suggesting it may need further optimization.

The LaBSE model emerged as the best-performing baseline, demonstrating strong multilingual handling and adaptability. However, challenges such as skewed keyword distributions led to overfitting, highlighting the need for balanced datasets and additional mitigation techniques.

Despite improvements in keyword classification, there was a slight decrease in sentiment classification accuracy with the fine-tuned models, indicating the complexity of multi-task learning. Future research should aim to optimize these adaptation techniques, address dataset biases, and enhance multi-task performance.

Acknowledgments

We thank our advisor Boshko Koloski, for his guidance through out the creation of the project.

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] Jože Bučar. Manually sentiment annotated slovenian news corpus SentiNews 1.0, 2017. Slovenian language resource repository CLARIN.SI.

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

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

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

[6] 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.

[7] 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.

Appendix

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 4. Best performing multi-label classification configurations.

Config	Epochs	Weight Decay	Number of Sentences	Scheduler	Learning Rate
1	1	0.0	20000	constantlr	0.0001
2	10	0.0	20000	constantlr	0.0001
3	10	0.0	20000	warmupconstant	0.0001
4	5	0.0	20000	warmuplinear	0.0001
5	5	0.0	20000	warmupcosine	0.0001

Table 5. Best performing multi-task classification configurations using the TSDAE adaptation technique.

Config	Epochs	Weight Decay	Number of Sentences	Scheduler	Learning Rate
1	1	0.01	20000	constantlr	0.00001
2	1	0.01	10000	constantlr	0.00001
3	1	0.01	70000	warmuplinear	2e-05
4	1	0.01	20000	warmuplinear	1e-5
5	5	0.01	50000	warmuplinear	1e-5

Table 6. Best performing multi-task classification configurations using the GPL adaptation technique.