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

Unsupervised Domain adaptation for Sentence Classification

Marko Možina, Peter Kosem, Aljaž Konec

Abstract

TBA

Keywords

Unsupervised Sentence Classification, Generative Pseudo Labeling, Transformer-based Denoising AutoEncoder

Advisors: Boshko Koloski

Introduction

Natural Language Processing (NLP) significantly benefits applications like sentiment analysis, spam detection, and topic categorization. However, the effectiveness of NLP models is often limited by the availability of labeled data, which can be scarce or expensive to obtain. Unsupervised domain adaptation techniques offer a solution to this problem by leveraging unlabeled data to enhance model performance.

In this report, we showcase adaptation of sentence transformer models to the specialized Slovenian domain SentiNews [**?**]. The dataset includes manually and automatically sentiment annotated sentences from Slovenian news articles The main adaptation techniques used are Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL). These methods refine the embedding space, making models more sensitive and accurate for specific domains, thereby improving sentence classification outcomes.

First we describe the Transformer-based Denoising AutoEncoder (TSDAE) [**?**], followed by Generative Pseudo Labeling (GPL) [**?**] method. Section presents the implementation details and evaluation of the adapted models. Finally, we discuss the results and outline future work in Section **??**.

Transformer-based Denoising AutoEncoder (TSDAE)

The core idea of TSDAE is to introduce noise to input sequences by deleting or swapping tokens (e.g., words). This corrupted input is then fed into the encoder component of the Transformer, whic produces a latent representation of the input. Decoder network, which also consists of transformer layers, then aims to reconstruct the original, clean input data from the latent representation. Below, we briefly explain the sequential process of TSDAE:

1. **Corruption:** The input data is corrupted with deleting a certain number of words, introducing variations and disturbances into the data. Adopting only deletion as the input noise and setting the deletion ratio to 0.6 performs best per [**?**].

2. **Encoding:** The corrupted input data is fed into the encoder, which consists of transformer layers. These layers transform the input data into a latent space representation called sentence vector, capturing essential features while filtering out noise.

3. **Decoding:** The latent representation obtained from the encoder is passed through the decoder, which aims to reconstruct the original, clean input data from the latent representation.

4. **Reconstruction:** The classifier token (CSL) embedding is used during reconstruction from token-level to sentence-level representation [**?**].

5. **Training:** The TSDAE optimizes its parameters by minimizing the reconstruction error between the denoised output generated by the decoder and the original, clean input data. This process occurs iteratively, allowing the model to learn effective denoising strategies.

By leveraging the Transformer architecture, TSDAEs can efficiently capture complex dependencies and patterns in the data, making them effective for denoising tasks across various domains, including natural language processing. Wang et al. [**?**] demonstrated that TSDAEs do not match the performance of supervised methods but remain valuable when data is unlabeled or difficult to obtain.

GPL

This section outlines the approach taken to adapt sentence-transformer models for improved sentence classification within specialized domains, leveraging the techniques of Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL).

Setup and Implementation

Model selection

Because the adaptaion is done on Slovenian text, we selected three sentence-transformer models that have been pre-trained on Slovenian data. The selected sentence transformer models are:

1. **bert-base-multilingual-uncased** [**?**]: A multilingual BERT model that has been pre-trained on a diverse range of languages, including Slovenian.
2. **all-MiniLM-L12-v2** [**?**]: A lightweight transformer model that has been pre-trained on a large corpus of text data.
3. **hate_speech_slo** [**?**]: A transformer model that has been fine-tuned on a Slovenian hate speech dataset.

TSDAE

For training data we randomly selected 100 000 sentences from the automatically annotated SentiNews dataset. The mannualy annotated sentences are in general more reliable and therefore we used them for testing. As mentioned before the sentences were corrupted by using a deletion ration of 60 % with a batch size of 16. The learning rate was set to 0.00001 and the models were trained for 50 epochs.

Generative pseudo labeling (GPL)

In the vast landscape of digital information, the ability to effectively process and classify text across diverse domains remains a paramount challenge in natural language processing. Traditional models often falter when applied outside their training domain due to the unique linguistic characteristics of new data sets. This gap highlights the urgent need for domain adaptation techniques capable of leveraging the wealth of unlabeled textual data prevalent in specialized fields. Generative Pseudo Labeling (GPL) emerges as a vital solution, offering a novel approach to utilize unlabeled data for enhancing model adaptability and performance in uncharted domains.

Generative Pseudo Labeling (GPL) is predicated on the innovative use of unlabeled data to improve model functionality in target domains. The GPL methodology unfolds in two pivotal stages:

1. **Pseudo Label Generation:** A pre-trained model, proficient in a related but distinct task, assigns provisional labels to unlabeled target domain data. These initial labels, derived from the model's pre-existing knowledge, serve as a foundational step for domain adaptation [**?**].
2. **Refinement through Generative Modeling:** Subsequently, the model undergoes a self-enhancement phase, refining its capabilities by learning from the data directly. This involves generative models that discern and adapt to the underlying patterns specific to the target domain, thereby aligning the model more closely with the target domain's characteristics [**?**].

Our project seeks to leverage GPL for the unsupervised domain adaptation of sentence-transformer models, aiming to bolster sentence classification accuracy within specialized domains. The application process is outlined as follows:

1. **Initial Model Training:** Employing a pre-trained sentence-transformer model, leveraging its extensive knowledge base for a preliminary understanding of the target domain [**?**].
2. **Pseudo Label Creation:** Generating pseudo labels for the Slovenian classification dataset (e.g., SentiNews) with the pre-trained model, bridging the model's knowledge from general to specific domains.
3. **Model Adaptation via GPL:** A generative model refines the sentence embeddings and classification efficacy of the sentence-transformer, emphasizing the adaptation to capture domain-specific nuances accurately [**?**].
4. **Iterative Refinement and Evaluation:** Through continuous refinement and evaluation, the model's performance is iteratively improved, ensuring its alignment with the project's goals.

Results

Sentence Classification Performance

To evaluate the performance of the adapted models, we trained a simple logistic regression classifier on the sentence embeddings generated by the models. The classifier was trained on a subset of 38 000 sentences from mannualy annotated SentiNews dataset, with an 80-20 train-test split. Figure 1 shows F1 scores and log loss for the base and TSDAE adapted models.

Model	F1 Score	Log Loss
bert-base-multiling.-uncased	0.606	0.756
gpl-bert-base-multiling.-uncased	TBA	TBA
hate_speech_slo	0.606	0.756
gpl-hate_speech_slo	TBA	TBA
all-MiniLM-L12-v2	0.561	0.820
gpl-MiniLM-L12-v2	TBA	TBA

Table 1. Performance metrics of base and GPL models.

Analysis

The results indicate a slight performance decrease for TSDAE models compared to their base counterparts across the models

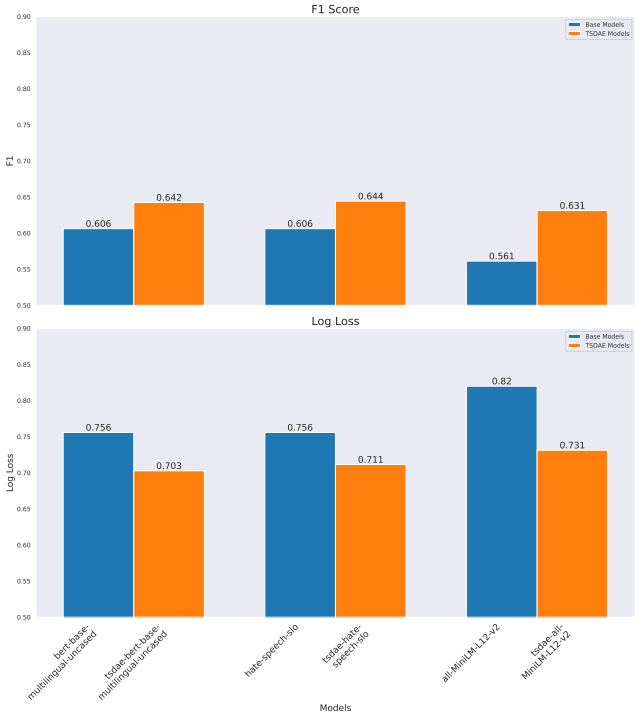

Figure 1. Performance metrics of base and TSDAE models.

tested. This could be attributed to several factors, including the complexity of the domain adaptation task, the characteristics of the dataset, and the inherent challenges in using unsupervised methods for domain-specific sentence classification. Additionally, the TSDAE models were only trained on 10,000 sentences, which may have contributed to their relatively poorer performance due to insufficient training data. Furthermore, the increase in log loss suggests that while TSDAE models may be refining the sentence embeddings, they might also be introducing aspects that do not correlate as effectively with the classification objectives as the original models.

Analysis for GPL results TBA.

Future Work

Further investigations will focus on optimizing the TSDAE training process, exploring different configurations of noise and corruption techniques, and increasing the dataset size to improve the robustness and accuracy of the models. Additionally, deeper analysis into the types of errors made by the TSDAE models may provide insights into their operational dynamics and potential areas for enhancement.

Future work in the domain of Generative Pseudo Labeling (GPL) for Slovenian text could focus on several avenues to enhance its effectiveness. Firstly, exploring novel techniques for generating pseudo labels specific to Slovenian language nuances could refine the embedding space further. Additionally, investigating methods to adapt GPL models for different text classification tasks within the Slovenian language domain could broaden its applicability. Furthermore, integrating semi-supervised learning approaches with GPL could leverage unlabeled data more effectively, potentially boosting classification performance. Finally, conducting thorough evaluations of GPL models on diverse Slovenian text corpora to assess their robustness and generalization capabilities would be crucial for practical deployment and widespread adoption.

References