FRI Natural language processing course 2024
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Unsupervised Domain adaptation for Sentence Classification

Marko Možina, Peter Kosem, Aljaž Konec

Abstract

In this study, we explore Unsupervised Domain Adaptation for Sentence Classification using Transformer-based Denoising Autoencoders (TSDAE) and Generative Pseudo Labeling (GPL). We utilized pretrained models on 100,000 sentences from the automatically annotated SentiNews dataset. Our findings indicate that TSDAE achieved superior performance compared to untrained models, demonstrating its effectiveness in capturing complex dependencies and improving classification accuracy in the absence of labeled data. Specifically, the MiniLM model's F1 score improved from 0.561 to 0.631, and log loss decreased from 0.820 to 0.731. However, applying GPL on top of TSDAE resulted in a slight decrease in performance, with F1 scores dropping by 0.5% to 2%. These results highlight the potential of TSDAE in unsupervised domain adaptation of Slovenian Language while suggesting that further refinement is needed for the combined TSDAE and GPL approach.

Keywords

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

Advisors: Boshko Koloski MSc, Researcher.

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 [1]. 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.

Very few domain adaptations have been done to a Slovenian dataset, therefore we first discuss Related Work in this field in other languages and then we describe the Transformer-based Denoising AutoEncoder (TSDAE) [2], followed by Generative Pseudo Labeling (GPL) [3] method. Section presents the implementation details and evaluation of the adapted models. Finally, a conclusion and future work is discussed.

Related Work

Unsupervised sentence embedding

Unsupervised sentence embedding learning has seen considerable advancements, driven by the need to generate robust and generalizable representations without extensive labeled data. Among the recent contributions to this field is the Transformer-based Denoising Autoencoder (TSDAE) introduced by Wang et al. (2021), which combines the strengths of transformers and denoising autoencoders. TSDAE introduces noise into input sequences and utilizes transformer layers for encoding and decoding, effectively capturing complex patterns in the data. Figure 2 showcases the TSDAE pipeline. This method has shown promising results in unsupervised domain adaptation by enhancing the robustness and accuracy of sentence embeddings.

Unsupervised domain adaptation

In recent years, unsupervised domain adaptation for dense retrieval has gained significant attention due to the growing need to adapt retrieval models to specific domains without requiring extensive labeled data. One notable contribution in this area is Generative Pseudo Labeling (GPL) that leverages generative models to create pseudo labels for unlabeled data, thereby enhancing the training process in semi-supervised and unsupervised learning scenarios. Reimers et al. (2019) and Wang et al. (2022) demonstrated that GPL could significantly

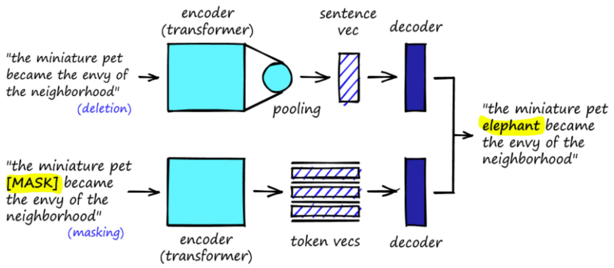
encoder
(transformer)
sentence
vec
decoder
"the miniature pet became the envy of the neighborhood"
(deletion)
pooling
"the miniature pet elephant became the envy of the neighborhood"
"the miniature pet [MASK] became the envy of the neighborhood"
(masking)
encoder
(transformer)
token vecs
decoder

Figure 1. TSDAE adaptation pipeline [4].

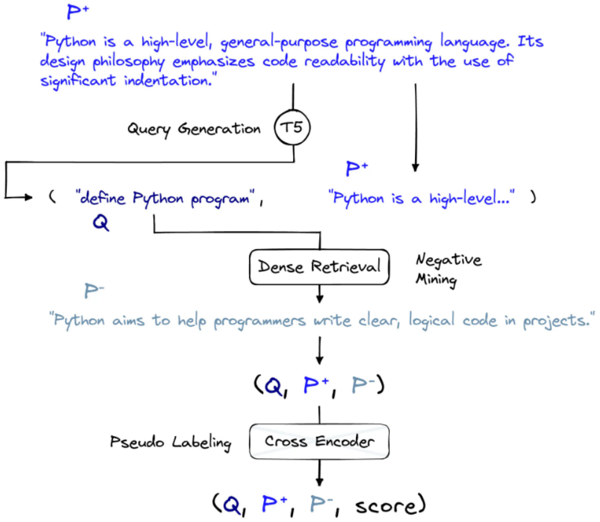
P+
"Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation."
Query Generation
T5
("define Python program",
Q
P+
"Python is a high-level...")
Dense Retrieval
Negative Mining
P-
"Python aims to help programmers write clear, logical code in projects."
(Q, P+, P-)
Pseudo Labeling
Cross Encoder
(Q, P+, P-, score)

Figure 2. Overview of the GPL process [5].

improve model performance by generating high-quality synthetic labels and iteratively refining the model's capabilities to adapt to target domains.

Application of Sentence Transformers

Sentence transformers have been widely adapted for various NLP tasks, including sentiment analysis, text classification, and translation. Their ability to generate meaningful sentence embeddings makes them ideal for domain adaptation. Several studies have successfully applied sentence transformers to low-resource languages and domain-specific datasets, highlighting their versatility and effectiveness.

Challenges and Future Directions

Despite advancements, domain adaptation in NLP faces challenges such as data scarcity, domain mismatch, and high computational costs. Current methods like TSDAE and GPL help, but innovative approaches are needed to better handle these issues. Future research should focus on efficient model architectures, advanced transfer learning, and comprehensive evaluation methods to enhance domain-specific adaptations and improve model performance.

Methods

Transformer-based Denoising AutoEncoder (TSDAE)

The core idea of TSDAE [2] 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].

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 [6].

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. [2] demonstrated that TSDAEs do not match the performance of supervised methods but remain valuable when data is unlabeled or difficult to obtain.

Generative Pseudo Labeling (GPL)

GPL leverages generative models to create pseudo labels for unlabeled data, enhancing the training process. By generating synthetic labels, GPL helps to expand and diversify the training dataset, improving model performance, especially in scenarios with limited labeled data. This approach combines the strengths of generative models and semi-supervised learning to achieve more robust and accurate predictions. Its 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 [7].

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 [8].

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 [7].

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 [8].

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.

By leveraging generative models, GPL can efficiently create high-quality pseudo labels for unlabeled data, enhancing the training process across various domains. Although GPL may not always match the performance of fully supervised methods, it remains valuable when labeled data is scarce or difficult to obtain. This approach is particularly effective in improving model performance in semi-supervised learning scenarios.

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** [9]: A multilingual BERT model that has been pre-trained on a diverse range of languages, including Slovenian.

2. **all-MiniLM-L12-v2** [10]: A lightweight transformer model that has been pre-trained on a large corpus of text data.

3. **hate_speech_slo** [11]: 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.

GPL

GPL employs a distinct strategy for domain adaptation. Initially, the sequences must be formatted according to the BeIR data structure and saved as a `jsonl` file. For training, we utilized the official GPL library [3]. The `train` method manages query generation, negative mining, and pseudo-labeling processes. We experimented with multiple variations, specifically 10,000, 20,000, 50,000, and 100,000 sentences, and conducted training with 10,000, 50,000, 100,000, and 140,000 GPL steps.

Results

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 manually annotated SentiNews dataset, with an 80-20 train-test split. Figure 3 shows F1 scores and log loss for base, TSDAE and TSDAE+GPL adapted models. All TSDAE models perform better than their base counterparts. MiniLM model showed the biggest improvement, with the F1 score increasing from 0.561 to 0.631 and Log loss decreasing from 0.820 to 0.731. This is due to the small size of the model, allowing it to adapt to a new domain faster and with less training examples. The other two models show a around a 4 % increase in F1 score and a 0.05 decrease in log loss.

Applying Generative Pseudo Labeling on top of the TSDAE models resulted in worse performance. All models see a 0.5% - 2% decrease in F1 score.

Table 1 shows the performance of the tested combinations of GPL and TSDAE models. Contrary to the TSDAE results, all GPL adapted models performed worse than their base counterparts. The all-MiniLM-L12-v2 model did not see an improvement in performance and gives the worst results in each of the tested combinations of results. Using more training sequences and larger models gave the best results.

Model Type	Adaptation	Training sequences	GPL steps		
		(thousands)	(thousands)	F1	Log-Loss
all-MiniLM-L12-v2	gpl	50	70	0.534	0.841
bert-base-multilingual-uncased	gpl	50	70	0.585	0.788
hate_speech_slo	gpl	50	70	0.577	0.79
all-MiniLM-L12-v2	tsdae + gpl	50	70	0.613	0.755
tsdae-bert-base-multilingual-uncased	tsdae + gpl	50	70	0.638	0.731
hate_speech_slo	tsdae + gpl	50	70	0.635	0.73
tsdae-all-MiniLM-L12-v2	tsdae + gpl	10	140	0.614	0.757
hate_speech_slo	tsdae + gpl	10	140	0.632	0.739
tsdae-bert-base-multilingual-uncased	tsdae + gpl	10	140	0.606	0.782
tsdae-all-MiniLM-L12-v2	tsdae + gpl	100	50	0.621	0.751
hate_speech_slo	tsdae + gpl	100	50	**0.64**	**0.725**

Table 1. Performance evaluation of tested combinations of GPL and TSDAE models.

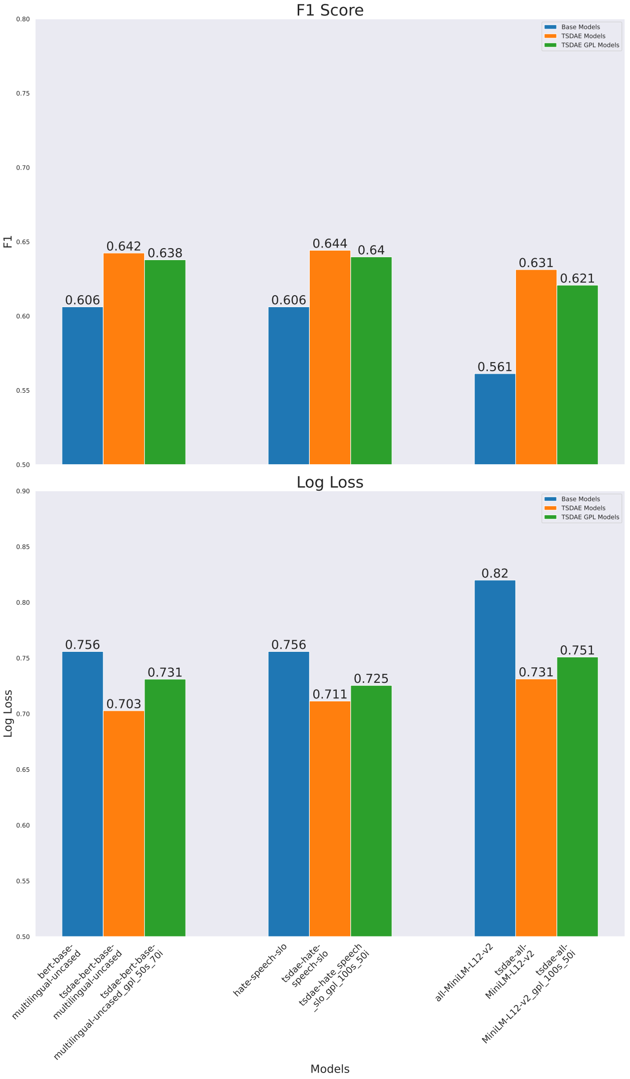

Figure 3. Performance comparison of base and TSDAE models.

Conclusion

It's possible that methodological issues impacted our comparison between TSDAE and GPL. Factors like hyperparameter tuning, data preprocessing, choice of evaluation metrics, experimental design, and implementation details could have influenced results. Addressing these factors and conducting further experiments may provide clearer insights into the performance differences between the two methods.

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

[1] Jože Bučar, Martin Žnidaršič, and Janez Povh. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919, 2018.

[2] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising autoencoder for unsupervised sentence embedding learning, 2021.

[3] 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*, 4 2021.

[4] Unsupervised training for sentence transformers — pinecone. https://www.pinecone.io/learn/series/nlp/unsupervised-training-sentence-transformers/?fbclid=IwAR0MCoGcYz83jnTtE0jshgq59h7HR4Jd0iwGoxts5UXoQWU-JDzQWNvMQiM. (Accessed on 05/24/2024).

[5] Domain adaptation with generative pseudo-labeling (gpl) — pinecone. https://www.pinecone.io/learn/series/nlp/gpl/#A-Simpler-Approach. (Accessed on 05/24/2024).

[6] Unsupervised training for sentence transformers, 2021. (Accessed on 03/21/2024).

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

[8] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States, July 2022. Association for Computational Linguistics.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[10] Hugging Face. all-minilm-l12-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2, 2023. Accessed: 2024-05-24.

[11] Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelicon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. Handling disagreement in hate speech modelling. In Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślezak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 681–695, Cham, 2022. Springer International Publishing.