Performance evaluation of the SloBERTa and XML-RoBERTa transformer models on the Slovenian news corpus SentiNews 1.0

Martin Domajnko
martin.domajnko@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

Jakob Kordež
jakob.kordez@student.um.si
Faculty of Electrical
Engineering and Computer Science,
University of Maribor
Koroška cesta 46
SI-2000 Maribor, Slovenia

ABSTRACT

Sentiment analysis, also called opinion mining, is a highly restricted natural language processing problem. This paper presents the use of existing SloBERTa and XML-RoBERTa models on the Slovenian news corpus SentiNews 1.0 and compares their performance. The results are further compared to the results achieved by the Multinomial Naive Bayes and Support Vector Machines methods used in the dataset paper. The trained models are also applied to data collected from the social media platform Reddit, in order to analyse the sentiment of posts and comments from the Slovenian community.

KEYWORDS

sentiment analysis, transformers, natural language processing, RoBERTa

1 INTRODUCTION

The rapid growth of digitally recorded opinion data over the past two decades is a key driver of the growing popularity of sentiment analysis. A lot of research is focused on collecting and analysing the data from review websites, forums and social media platforms like Twitter [1, 16], Internet Movie Database (IMDB) [9] and Amazon [11]. The data is collected with the help of web scraping tools and the social media platforms's own public API services.

Sentiment analysis research has been mainly carried out at three levels of granularity: document level, sentence level, and aspect level [7]. The problem can be approached as a binary classification problem, where the text is classified as positive or negative, or as a multi-class classification problem, where the text is classified as positive, negative, or neutral. The latter can also use a different set of three or more classes.

Firstly, a brief overview is given of the different approaches to sentiment analysis and a more exhaustive description of the most popular methods currently. This is followed by an outline of the dataset we used to train and evaluate the selected models and an examination of the data we scraped from the social media platform Reddit. Section 4 describes the training and evaluation pipeline. Results and findings are presented in Section 5 and the paper conludes in Section 6 with possible improvements.

2 RELATED WORK

Reviews of products or services and social media posts are the primary targets for sentiment classification. The employed techniques can be divided into three groups according to the methods they use. The machine learning approach uses machine learning models in combination with linguistic features, and can be broken down into supervised and unsupervised learning methods. The lexicon-based approach utilizes pre-prepared sentiment lexicons and is also divided into dictionary-based methods and corpus-based methods. The final approach is the hybrid approach, which combines the aforementioned methods [12].

Using computational methods, sentiment analysis systematically analyses people's expressed subjective information, such as opinions and emotions, towards different entities. The current state-of-the-art performance is realized with deep learning models using the means of self-attention, also known as transformers [18]. The concept was first presented by the combined team from Google Brain and Google Research in their 2017 paper "Attention is all you need" [18]. The Transformer model solved the recurrent neural networks and long short-term memory networks biggest constraint, that of sequential computation. The model doesn't rely on recurrence, but it insted relies entirely on an attention mechanism to draw global dependecies between input and output [18]. The model achieves faster computation times, because it allows for significantly more parallelization.

The most widely used language representation model is BERT, which stands for Bidirectional Encoder Representations from Transformers. It leverages the benefits of transformers and extends them with the ability to train the models on large unlabelled datasets and then fine-tune them, with just one additional output layer, on a wide range of downstream tasks [5]. One downside of these models is that, because of their size, they are hard to run in constrained computational environments. Authors in [15] proposed a distilled version of the BERT model called DistilBERT, which reduces the model's size by 40% and increases its speed by 60% while still achieving 97% of its performance. Further research on the BERT model has shown it to be considerably under trained [8]. The model RoBERTa, which stands for Robustly optimized BERT approach, proposed some modifications to the pre-training procedure that improved end-task performance and achieved state-of-the-art results on GLUE [19], SQuAD [14] and RACE [6] datasets [8]. Another important mention is the transformer-based multilingual masked language model XLM-R, which was pre-trained on text in 100 languages [4]. In our work, we compare it against the monolingual Slovene RoBERTa model [17] in two and three class sentiment analysis tasks.

Table 1: Number of negative, neutral and positive examples in the dataset based on the level of granularity

Level of	Sentiment			Total
granularity	Negative	Neutral	Positive	
Sentence	26.74%	56.98%	16.28%	168,899
Paragraph	26.36%	57.38%	16.26%	89,999
Document	32.00%	52.03%	15.97%	10,427

3 USED DATA

3.1 Dataset

For the training and evaluation of the models, we used the manually sentiment annotated Slovenian news corpus SentiNews 1.0 [2]. The data is sentiment annotated on three levels of granularity: sentence, paragraph and document level [2]. Our models were only trained on the sentence and paragraph levels. The same corpus is also presented in [3], where it was used to train two classifiers (Multinomial Naive Bayes and Support Vector Machines). We also compare the performance of our models to the results given in [3].

We split the corpus randomly into training, validation and testing sets with the sklearn Python library. The training set contained 90% of the data, while each of the other two contained 5%. As seen in Table 1, the data is heavily imbalanced and favours the neutral class.

3.2 Scraped data

We decided to scrape the r/Slovenia subreddit because the majority of the posts there were in Slovene. We accomplished this with a Python 3 script using the `praw`[1] package, which is a Reddit API wrapper. Because we wanted to scrape the whole subreddit, we also used the package `psaw`[2] which in combination with `praw` enabled us to fetch more than only 1000 posts that we were limited to before.

Using the first script, we were able to obtain the IDs and flairs of all the posts, group them by their flair (category), and save them locally. Our second script then enabled us to individually fetch posts by their IDs. We decided to fetch the post title, body text, score, upvote ratio, a timestamp of when it was created and the comments which we flattened from a tree structure to a list. Each comment has a numerical score and a text entry. After we removed all posts that were simply links to other posts and filtered the remaining posts, we were left with 14,000 posts which combined contained nearly 240,000 comments.

Not all posts and comments were in Slovene, some were in English. Most of the English posts were foreign people asking questions, so we decided to filter only Slovene posts and comments. We accomplished this with the help of N-Gram-Based text categorization with a Slovene and an English corpus. After filtering, we were left with 9,000 posts and 174,000 comments.

4 METHOD

Transformer models follow an encoder-decoder structure. For them to be able to process the text, it is first transformed into numerical vectors, also called word embeddings. Since the models don't use any recurrence, the word embeddings are also combined with positional encodings, which present information about the relative or absolute position of the tokens in the sequence [18]. Our work fine-tuned the XLM-RoBERTa (XLM-R) [4] and SloBERTa [17] models for Sentiment Analysis. PyTorch [13] based implementations of the models from the open-source Transformers library [20] were used. The models and tokenizers were already pre-trained on large datasets.

The architecture and training approach used in the SloBERTa model [17] is the same architecture as the RoBERTa base model [8], but it uses a different tokenization technique. For this, a Sentence Piece[3] model was trained, which splits the text into tokens and encodes it into subword byte-pair encodings. The model has a vocabulary size of 32000 subword tokens [17]. The model is closely related to the French monolingual model CamemBERT [10]. The pre-trained SloBERTa model we used, is the second version of the model, which was trained for 200000 updates in total [17].

XLM-R [4] is a multilingual version of the RoBERTa [8] model. Instead of using language-specific preprocessing and byte-pair encoding, the model uses a Sentence Piece model trained on raw text data for all languages. The pre-trained XLM-R model we used, was trained for 1.5 Million updates on five-hundred 32GB Nvidia V100 GPUs with a batch size of 8192 and has a vocabulary size of 250000 [4].

5 RESULTS

5.1 Experimental setting

The Kaggle environment with an Nvidia Tesla P100 GPU was used for training the models. We used a batch size of 32 and a learning rate of 6×10^{-7} for the SloBERTa model [17] and a batch size of 8 and a learning rate of 1×10^{-7} for the XLM-R model [4]. The hyperparameters were selected empirically. Additionally, to reduce the effects of class imbalance in the training data set, each model was supplied with precomputed class weights. The models were trained for a total of 20 epochs.

5.2 Model performance

The models, trained on the dataset annotated on the paragraph level of granularity, were evaluated on binary (positive and negative) and multi-class (positive, negative, and neutral) sentiment analysis tasks. The results in Table 2 show that on the binary classification task, both the SloBERTa model and XLM-R model achieve similar results with an accuracy above 90%. On the other hand, the multi-class classification task results show that the XLM-R model performs better, with an accuracy of 70.49%, compared to the accuracy of the SLoBERTa model of 66.04%. Additionally, the SloBERTa model trained on the dataset annotated on the sentence level of granularity, was also evaluated on binary and multi-class sentiment analysis tasks. The model's performance was compared to the SVM and NBM models presented in paper [3]. The results are collected in Table 3 and show that the much simpler NVM and NBM models outperform the SloBERTa model on both binary and multi-class classification tasks. Additionally, all the models achieve an accuracy

[1] https://github.com/praw-dev/praw

[2] https://github.com/dmarx/psaw

[3] https://github.com/google/sentencepiece

Table 2: Comparison of results on the dataset annotated on the paragraph level of granularity, for the SloBERTa and XLM-R models, on binary (positive and negative) and multi-class (positive, negative, and neutral) sentiment analysis tasks. The first two columns contain the model name and number of classes. The other four columns present the model accuracy, precision, recall and F1 norm metrics.

Model	No. of classes	Accuracy	Recall	Precision	F1
SloBERTa	2	91.19	90.91	90.67	90.79
XLM-R	2	91.29	90.83	90.89	90.86
SloBERTa	3	66.04	72.19	64.74	65.70
XLM-R	3	70.49	68.93	67.84	68.35

Table 3: Comparison of results on the dataset annotated on the sentence level of granularity, for the SloBERTa model and the two models (SVM and NBM) presented in paper [3], on binary (positive and negative) and multi-class (positive, negative, and neutral) sentiment analysis tasks. The first two columns contain the model name and number of classes. The other two columns present the model accuracy and F1 norm metrics.

Model	No. of classes	Accuracy	F1
SloBERTa	2	90.40	89.87
SVM	2	93.10	94.86
NBM	2	95.21	96.38
SloBERTa	3	65.47	64.64
SVM	3	73.10	55.35
NBM	3	66.46	61.20

above 90% on the binary task and an accuracy above 65% on the multi-class task.

5.3 Reddit analysis

The goal of this experiment was to analyse the sentiment of posts and comments made in the Slovenian community on the social media platform Reddit. We also looked for a potential correlation between the score, flair, and sentiment of the analysed posts and comments.

In the process of fine-tuning, a model trained to perform on a given task is tweaked to perform on a second similar task. We used the SloBERTa and XLM-R models, which were fine-tuned for binary and multi-class sentiment analysis tasks on the dataset annotated at the paragraph level of granularity, to predict the sentiment of the 9,000 posts and 170,000 comments scraped from the social media platform Reddit.

Table 4 summarizes the sentiment classification distribution of posts and comments. We can see that the multi-class SloBERTa model classifies the posts and comments evenly between the three sentiments, while the XLM-R model classifies more posts and comments as neutral.

Table 4: Classification results for posts and comments, made by the SloBERTa and XLM-R models trained on the dataset annotated on the paragraph level of granularity.

Model	No. of classes	Negative	Neutral	Positive
SloBERTa	2	48.66%	/	51.34%
XLM-R	2	56.53%	/	43.47%
SloBERTa	3	32.66%	37.80%	29.54%
XLM-R	3	22.08%	61.22%	16.70%

We found that posts and comments with more downvotes than upvotes tend to have a negative sentiment rather than a positive one. Furthermore, longer comments and comments with a score above 250, are more likely to be classified as negative.

When we grouped posts by their flair, we found that posts tagged with the flair “Question” had a much larger number of neutral posts. The flairs “News” and “Article” had an equal number of posts with positive and negative sentiments, but their comments were heavily leaning to the negative side. We also found that posts and comments tagged with the flair “Discussion” were more negative than positive.

6 CONCLUSION

In this study, we fine-tuned two transformer models, SloBERTa and XLM-R, for binary and multi-class sentiment analysis tasks on the Slovenian news corpus SentiNews 1.0. Both models achieved similar results on the binary classification task, while on the multi-class classification task, the XLM-R model performed better. Additional comparison of the SloBERTa model with the NVM and NBM models has shown that the transformer model achieves slightly worse results on both binary and multi-class classification tasks. The trained models were also applied on data scraped from the social media platform Reddit. The results have shown that the XLM-R model is more likely to classify posts and comments as neutral, while the SloBERTa model classifies all classes evenly.

REFERENCES

[1] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. https://doi.org/10.48550/ARXIV.2010.12421

[2] Jože Bučar. 2017. Manually sentiment annotated Slovenian news corpus SentiNews 1.0. http://hdl.handle.net/11356/1110 Slovenian language resource repository CLARIN.SI.

[3] Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation* 52, 3 (Feb. 2018), 895–919. https://doi.org/10.1007/s10579-018-9413-3

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.48550/ARXIV.1911.02116

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. https://doi.org/10.48550/ARXIV.1704.04683

[7] B. Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions.* 1–367 pages. https://doi.org/10.1017/CBO9781139084789

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa:

A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[9] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 142–150. https://aclanthology.org/P11-1015

[10] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suá rez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645

[11] Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China) *(RecSys '13)*. Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163

[12] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5, 4 (2014), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[14] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. https://doi.org/10.48550/ARXIV.1606.05250

[15] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

[16] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

[17] Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. http://hdl.handle.net/11356/1397 Slovenian language resource repository CLARIN.SI.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[19] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. https://doi.org/10.48550/ARXIV.1804.07461

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6