MultiEmo: Language-agnostic Sentiment Analysis [⋆]

Piotr Miłkowski, Marcin Gruza, Przemysław Kazienko, Joanna Szołomicka,
Stanisław Woźniak, and Jan Kocoń

Department of Artificial Intelligence, Wrocław University of Science and Technology,
Wrocław, Poland piotr.milkowski@pwr.edu.pl

Abstract. We developed and validated a language-agnostic method for sentiment analysis. Cross-language experiments carried out on the new MultiEmo dataset with texts in 11 languages proved that LaBSE embeddings with an additional attention layer implemented in the BiLSTM architecture outperformed other methods in most cases.

Keywords: cross-language NLP · sentiment analysis · language-agnostic representation · LASER · LaBSE · BiLSTM · opinion mining · MultiEmo

1 Introduction

Two of the most important and applicable topics in natural language processing (NLP), in particular in opinion mining, include sentiment analysis [1–3] and emotion recognition [4], [5]. Recently, more and more online comments are expressed in different natural languages. Consequently, there is a growing interest in new methods for sentiment analysis that are language-independent. For that purpose, appropriate language-agnostic models (embeddings) may be utilized.

In this paper, we developed and validated three language-agnostic methods for sentiment analysis: one based on the LASER model [6] and two on LaBSE [7], see Sec. 3. The latter was used in its basic version ($LaBSE_b$) and with additional attention layer ($LaBSE_a$). All of them were implemented within the bidirectional LSTM architecture (biLSTM). The experiments were performed on our new benchmark MultiEmo dataset, which is an extension of MultiEmo-Test 1.0 [8]. In the latter, only test texts were translated into other languages, whereas the MultiEmo data proposed here is fully multilingual. As the experiments revealed that LaBSE with the additional attention layer ($LaBSE_a$) performs best (Sec. 5), it was exploited in the MultiEmo web service for language-agnostic sentiment analysis: https://ws.clarin-pl.eu/multiemo. All results presented in this paper are downloadable: the MultiEmo dataset at https://clarin-pl.eu/dspace/handle/11321/798 and source codes at https://github.com/CLARIN-PL/multiemo.

[⋆] This work was partially supported by the National Science Centre, Poland, project no. 2020/37/B/ST6/03806; by the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN - Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10

2 Related work

Recently, in the domain of sentiment analysis, most research relies on effective solutions based on deep neural networks. The currently considered state-of-the-art apply Recurrent Neural Networks and Transformers, such as BiLSTM/CNN [9,10], BERT [10, 11], or RoBERTa [12, 13]. The idea of knowledge transfer between domains, document types, and user biases in the context of social media was discussed in [14]. However, *Language-agnostic Sentiment Analysis* is a less considered issue. This problem goes beyond the classical approaches that rely only on a single, resource-rich language, commonly English, and focuses on other languages.

In our previous work [8], we analyzed the task of cross-language sentiment analysis. In particular, we applied vector representations not depending on a particular language directly [6], hence, transferring knowledge from one language to another appears to be quite efficient. We also proposed a benchmark dataset containing test files translated into 8 different languages. However, we have not exploited state-of-the-art methods. Therefore, the dataset published in this study was a base for our further experiments.

Another interesting approach is zero-shot learning investigated by the Slovenian CLARIN-SI team in [15]. They used it for news sentiment classification. Given the annotated dataset of positive, neutral, and negative news in Slovenia, their goal was to develop a news classification system that not only assigns sentiment categories to Slovenian news, but also to news in another language, without any additional training data. Their system was based on the multilingual BERT model [11]. At the same time, they tested different methods of processing long documents and proposed a new BERT model using emotional enrichment technology as an intermediate training step. They also evaluated the zero-sample cross-language ability of their system on the new news sentiment test set in Croatian. Due to their work, their cross-language approach is also superior to most classifiers to a large extent, and all settings without emotional richness in pretraining.

Most of the most advanced sentiment classification methods are based on supervised learning algorithms that require large amounts of manually labeled data. However, annotated resources are often differently unbalanced and of little quantity among different languages. Cross-lingual sentiment classification solves this problem by using knowledge from high-resource languages to low-resource ones. In [16], an attention-based bilingual representation learning model was proposed. It can learn the distributed semantics of documents in the source language and the target language. In each language, the authors use long short-term memory (LSTM) networks to model documents, which have proven to be very effective for word sequences. At the same time, the hierarchical attention mechanism of bilingual LSTM network (BiLSTM) was proposed. The sentence-level attention model learns which sentences in the document are more important to determine the overall sentiment, while the word-level attention model learns which words in each sentence are decisive. The proposed model achieved good results on the benchmark dataset while being trained on English and evaluated on Chinese.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10

We tested LASER and LaBSE embeddings on multiple NLP tasks, including primarily studies on sentiment. In this paper, however, we go further and investigate them in the multiple language environment.

3 Language-agnostic models combined with LSTM

LASER Facebook Research prepared the LASER method for representing a text's sentences as a list of 1024 items long vectors supporting 93 languages [6]. It uses a common space for embeddings of sentences in any language. A single model generates embeddings for all languages. It is necessary to specify language of the input text, because sentence tokenization (dividing a document into a list of sentences) is language specific.

LaBSE Google AI team modified multilingual BERT (M-BERT) [17] to produce language-agnostic sentence embeddings (LaBSE) for 109 languages [7]. Their work produced a state-of-the-art mask language model, allowing task-specific fine-tuning what allows to achieve the best results by one's models across all supported languages. They trained the model with the usage of a translation ranking task. It is based on bidirectional dual encoders that results in a robust combination of masked language model (MLM) and translation language model (TLM) [18].

The LaBSE model prepared during the study and subsequently released increased the average dual-text retrieval accuracy for 112 languages to 83.7%. Note that if it is compared with the 65.5% accuracy achieved by LASER on the same benchmark Tatoeba corpus, it opens up new research directions in many tasks. For example, the potential application of LaBSE includes mining parallel text from the web, however, we want to test it on sentiment analysis.

The dual-encoder architecture [19] with Additive Margin Softmax [20] essentially uses parallel encoders to encode two sequences and then to obtain a compatibility score between both encodings using a dot product. The model uses MLM and TLM pretraining to train on 17 billion single sentences and 6 billion bilingual sentence pairs. The output model is quite effective even on low-resource languages where no data is available during the training. Compared to LASER, the combination of the two makes LaBSE perform better in most cases, see e.g. [21].

Performance metrics of LaBSE have been outstanding. According to our observation, LaBSE takes less than 500ms to generate embeddings for a sentence that is close to 20 words long. It is based on Transformer model with BERT-like architecture and pretrained on over 500k vocabulary.

In this work, we propose two LaBSE-based methods: (1) the basic usage of the LaBSE output containing plain embeddings ($LaBSE_b$) and (2) embeddings amended with the attention mask ($LaBSE_a$). The former is the classical approach often used to obtain text embeddings from transformer type models. Our attention-based variant first retrieves the token representations and then an attention mask by resizing it to match the size of the embedding output. Next, both matrices are multiplied with each other and summed up. In a later step, the summation and clamp of the attention mask are performed. The final product of this process is

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10

the division of the sum of embeddings by the prepared sum of the attention mask. This gives us an averaged embedding of tokens from the last output layer enriched with an indication of which tokens contain the most key information.

BiLSTM architecture A commonly used neural network with sentence embeddings is bidirectional long short-term memory (BiLSTM). A layer of such type learns bidirectional long-term dependencies between time steps of time series or sequence data. These dependencies can be useful when you want the network to learn from the complete time series at each time step. Our model operates on text's sentences encoded with LASER or LaBSE model. Overall, our deep architecture developed for the task of sentiment analysis consists of the following layers:

- The Gaussian noise layer with a standard deviation of 0.01 accepts input shapes of up to N sentences, and the vector matrix of each sentence is 1024, so the overall input shape is (N, 1024);
- a bidirectional layer with LSTM instances consisting of 1,024 hidden units, using hyperbolic tangent activation method;
- a dropout layer with a rate equal to 0.2;
- a dense layer with softmax activation (normalised exponential function) with 4 outputs representing probability of class occurrence for 4 output classes.

4 Experimental setup

4.1 Pipeline

Model training and evaluation were done in the following stages: (1) perform training on 80% of data and validation on 10%; (2) train the model until the loss function value stops decreasing for 25 epochs; keep the lowest achieved value of loss; (3) evaluate the trained model using the test part of data – the remaining 10%. All experiments were repeated 30 times so that strong statistical tests could be performed. This removed the amount of uncertainty caused by the randomness of the neural network model learning process. If the difference between the results in our statistical tests was $p < 5\%$, they were treated as insignificantly different.

4.2 MultiEmo dataset

We created a new kind of dataset for sentiment analysis tasks – PolEmo 2.0 [10]. Each sentence as well as the entire document are labelled with one out of the four following sentiment classes: (1) ***P***: positive; (2) ***0***: neutral; (3) ***N***: negative; (4) ***AMB***: ambivalent, i.e., there are both positive and negative aspects in the text that are balanced in terms of relevance. In all further experiments, we exploited only the labels assigned to the entire text – the document level processing. The whole MultiEmo corpus in Polish contains over 40K sentences. Since each text and sentence was manually annotated with sentiment in the 2+1 scheme, we received in total over 150K annotations. A high value of Positive Specific Agreement (PSA) [22] equal to 0.91 for texts and 0.88 for sentences was achieved.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10

Additionally, the whole corpus was machine translated into different languages using DeepL (`https://www.deepl.com/translator`), what resulted in a new MultiEmo dataset. It provides an opportunity to train and test the model in any out of 11 languages: Polish (origin), English, Chinese, Italian, Japanese, Russian, German, Spanish, French, Dutch and Portuguese. The comprehensive profile of the MultiEmo dataset is presented in Tab. 1. Only the mixed-domain corpus was exploited in the experiments described in Sec. 4.3 and 5, see the last row in Tab. 1.

Type	**Domain**	**Train**	**Dev**	**Test**	**SUM**	**Average length [chars]**
Mixed-domain texts (all domains)	Class *P* – positive	1,824	236	227	2,287	648
	Class *0* – neutral	971	128	118	1,217	854
	Class *N* – negative	2,469	304	339	3,112	817
	Class *AMB* - ambivalent	1,309	155	136	1,600	707
	All classes	**6,573**	**823**	**820**	**8,216**	**754**

Table 1: The number of texts in the train/dev/test set of the MultiEmo corpus. The average length is calculated for the entire set (SUM).

4.3 Scenarios

To validate the quality of the models, we used three research scenarios, differing in the language of the texts used to train and test the models:

- **Any->Same** – the model is both trained and tested on texts in one chosen language (e.g. Polish-Polish, English-English).
- **PL->Any** – the model is trained only on Polish texts and tested on docs translated to any other language (e.g. Polish-English, Polish-Chinese).
- **Any->PL** - the model is trained on texts in any language and tested only on Polish texts (e.g. English-Polish, Chinese-Polish, Dutch-Polish).

All scenarios use the same train-validation-test split, Tab. 1, which ensures that the model will not be trained and tested on the same translated texts.

5 Experimental results

The results for the same language training and testing on the MultiEmo dataset (all domains mixed), which is the first scenario described in Sec. 4.3, prove that $LaBSE_a$ is better in almost all cases. There are 5 situations when $LaBSE_b$ was insignificantly better than $LaBSE_a$. It happened in English (positive and negative labels), French (positive and neutral), and Italian (neutral).

In the second scenario, the training was carried out on Polish data and testing on other languages. $LaBSE_a$ is almost always statistically better than the other models. There are only eight cases out of all 88, in which $LaBSE_b$ was insignificantly better than $LaBSE_a$. There is also one situation (Portuguese: F1$_{samples}$) where $LaBSE_b$ is insignificantly worse than $LaBSE_a$. The results aggregated over all languages separately for each of the three considered models are shown in Fig. 1a for the LASER language model, in Fig. 1b for basic LaBSE ($LaBSE_b$), and in Fig. 1c for LaBSE with the custom mean pooling ($LaBSE_a$).

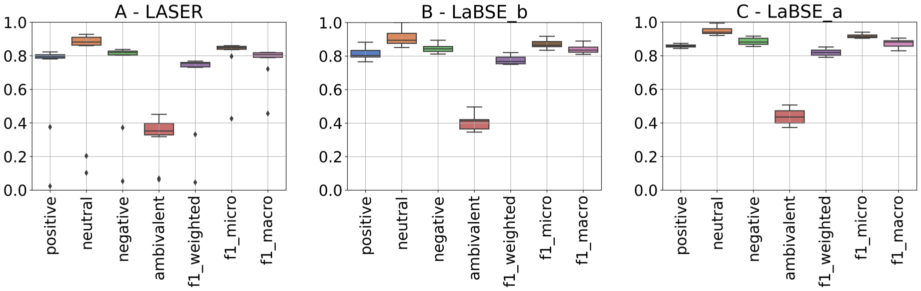

Fig. 1: Distribution of F1 scores for models learned on Polish texts and evaluated on all languages from the MultiEmo dataset (PL->Any scenario) aggregated over all test languages. (**A**) – for the LASER embeddings; (**B**) – for the basic $LaBSE_b$ embeddings; (**C**) – for the LaBSE with attention, i.e. $LaBSE_a$ embeddings

In the third scenario, the classifier was trained on different languages but testing was performed on Polish texts only. Similarly to the previous scenarios, $LaBSE_a$ outperforms $LaBSE_b$ and LASER language models. In all scenarios, the results for the *ambivalent* class are worse by about 40%-50% than for *negative* or *positive* class meaning some documents are more controversial than others. Rather, we should consider applying personalized reasoning to them [4, 5, 23, 24]. Also, the *neutral* class is poorly classified, especially for LASER and non-Latin languages (Chinese, Japanese, Russian). $LaBSE_a$ in the second scenario overcomes this problem revealing the superiority of language-agnostic solutions over language-specific ones. Languages using Latin alphabet perform almost the same.

6 Conclusion and Future Work

Language-agnostic embedding models can successfully provide valuable information for the classification of sentiment polarization. The experiments were carried out on the new multilingual MultiEmo dataset. They proved that language-agnostic representations are efficient. The best results were obtained for the LaBSE embeddings with an additional attention layer ($LaBSE_a$) and this solution was implemented in our online service. Performance for ambivalent documents may be unsatisfactory and demands other, e.g., personalized solutions. This will be investigated in future work.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10

References

1. F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review*.
2. Ł. Augustyniak, P. Szymański, T. Kajdanowicz, and P. Kazienko, "Fast and accurate - improving lexicon-based sentiment classification with an ensemble methods."
3. R. Bartusiak, L. Augustyniak, T. Kajdanowicz, and P. Kazienko, "Sentiment analysis for polish using transfer learning approach," in *ENIC 2015*.
4. P. Miłkowski, M. Gruza, K. Kanclerz, P. Kazienko, D. Grimling, and J. Kocon, "Personal bias in prediction of emotions elicited by textual opinions," in *ACL-IJCNLP 2021: Student Research Workshop*. ACL, 2021, pp. 248–259.
5. J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, and P. Kazienko, "Learning personal human biases and representations for subjective tasks in natural language processing," in *ICDM 2021*. IEEE, 2021, pp. 1168–1173.
6. M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the ACL*.
7. F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," *arXiv preprint arXiv:2007.01852*, 2020.
8. K. Kanclerz, P. Miłkowski, and J. Kocoń, "Cross-lingual deep neural transfer learning in sentiment analysis," *Procedia Computer Science*, vol. 176, pp. 128–137, 2020.
9. T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using bilstm-crf and cnn," *Expert Systems with Applications*.
10. J. Kocoń, P. Miłkowski, and M. Zaśko-Zielińska, "Multi-level sentiment analysis of polemo 2.0: Extended corpus of multi-domain consumer reviews," in *CoNLL'19*.
11. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding."
12. Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach."
13. P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik, "Klej: comprehensive benchmark for polish language understanding," *arXiv preprint arXiv:2005.00630*, 2020.
14. P. H. Calais Guerra, A. Veloso, W. Meira Jr, and V. Almeida, "From bias to opinion: a transfer-learning approach to real-time sentiment analysis," in *ACM SIGKDD'2011*.
15. A. Pelicon, M. Pranjić, D. Miljković, B. Škrlj, and S. Pollak, "Zero-shot learning for cross-lingual news sentiment classif." *Applied Sciences*, vol. 10, no. 17, p. 5993, 2020.
16. X. Zhou, X. Wan, and J. Xiao, "Attention-based lstm network for cross-lingual sentiment classification," in *EMNLP'16*, 2016, pp. 247–256.
17. T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" in *Proceedings of the 57th Annual Meeting of the ACL*, 2019, pp. 4996–5001.
18. L. Shen, J. Xu, and R. Weischedel, "A new string-to-dependency machine translation algorithm with a target dependency language model," in *ACL-08: HLT*.
19. M. Guo *et al.*, "Effective parallel corpus mining using bilingual sentence embeddings."
20. Y. Yang *et al.*, "Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax," *arXiv preprint arXiv:1902.08564*, 2019.
21. K. Gawron, M. Pogoda, N. Ropiak, M. Swędrowski, and J. Kocoń, "Deep neural language-agnostic multi-task text classifier," in *ICDM'21*. IEEE, 2021, pp. 136–142.
22. G. Hripcsak and A. S. Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *JAMIA*, vol. 12, no. 3, pp. 296–298, 2005.
23. J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko, "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach," *Information Processing & Management*, vol. 58, no. 5, p. 102643, 2021.
24. K. Kanclerz, A. Figas, M. Gruza, T. Kajdanowicz, J. Kocoń, D. Puchalska, and P. Kazienko, "Controversy and conformity: from generalized to personalized aggressiveness detection," in *ACL-IJCNLP 2021*. ACL, 2021, pp. 5915–5926.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08754-7_10