Improving Sentiment Classification in Slovak Language

Samuel Pecar, Marian Simko, Maria Bielikova
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovicova 2, 842 16 Bratislava, Slovakia
{samuel.pecar, marian.simko, maria.bielikova}@stuba.sk

Abstract

Using different neural network architectures is widely spread for many different NLP tasks. Unfortunately, most of the research is performed and evaluated only in English language and minor languages are often omitted. We believe using similar architectures for other languages can show interesting results. In this paper, we present our study on methods for improving sentiment classification in Slovak language. We performed several experiments for two different datasets, one containing customer reviews, the other one general Twitter posts. We show comparison of performance of different neural network architectures and also different word representations. We show that another improvement can be achieved by using a model ensemble. We performed experiments utilizing different methods of model ensemble. Our proposed models achieved better results than previous models for both datasets. Our experiments showed also other potential research areas.

1 Introduction and Related Works

Amount of text data produced by users in the world has grown rapidly in recent years. On the Web, users produce text using different platforms, such as social networks or portals aggregating customer reviews. Most of the produced text can be considered as opinionated. There is a significant need for utilization of natural language processing tasks, such as sentiment analysis or other connected tasks – emotion recognition, stance detection, etc.

Sentiment analysis can be viewed as one of the most common and widespread tasks in natural language processing. Recent advancements in neural networks allowed further research also for minor non-English languages. In recent years, there have been several studies researching sentiment classification of multiple Slavic languages, such as Czech (Habernal et al., 2014; Steinberger et al., 2014), Croatian (Rotim and Šnajder, 2017), Lithuanian (Kapočiutė-Dzikienė et al., 2013), Russian (Chetviorkin and Loukachevitch, 2013), and Slovak (Krchnavy and Simko, 2017; Pecar et al., 2018). Interesting study was also proposed by Mozetič et al. (Mozetič et al., 2016), where authors studied the role of human annotators for sentiment classification and provided also datasets for sentiment analysis of Twitter posts for multiple languages including some Slavic languages.

Whereas state-of-the-art methods widely employ different neural model architectures, such as the attention mechanism (Wang et al., 2016) or model ensemble techniques (Araque et al., 2017), recent research in sentiment analysis in Slavic languages still employs more traditional machine learning methods, mostly Support Vector Machines (SVM). We suppose this can be cause due to low availability of larger annotated datasets for Slavic languages, ones that are quite common for English or other major languages.

We see as an essential for further improvement of sentiment classification employing different techniques of transfer learning, especially using different pre-trained word representations on large text corpora. In recent years, there have been introduced many new methods for word representations, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018) or ULM-FIT (Howard and Ruder, 2018). Unfortunately, most of these pre-trained word representations are only available for English language and further training requires a significant amount of hardware resources and extensive text corpora. On the other hand, there have been recently introduced also word representations for other languages, such as pre-trained ELMo word representations (Che et al., 2018; Fares et al., 2017) or fastText (Grave et al.,

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 114–119,
Florence, Italy, 2 August 2019.

2018) for many different languages.

In this paper, we discuss possible methods for improving sentiment classification for Slovak language by using state-of-the-art methods. Our main contribution is employment of different neural model architectures for sentiment classification in Slovak. We also provide a study on how each block of architecture can contribute to overall sentiment classification.

2 Model

We believe that application of different neural network architectures can bring significant improvements of results. For our study, we consider employing several such architectures. A general architecture is shown in Figure 1 (Pecar et al., 2019). As shown in the figure, we consider four main block of this architecture, which are either variable or permanent. The last layer (linear decoder) is followed by logarithmic soft-max activation function to obtain final model predictions.

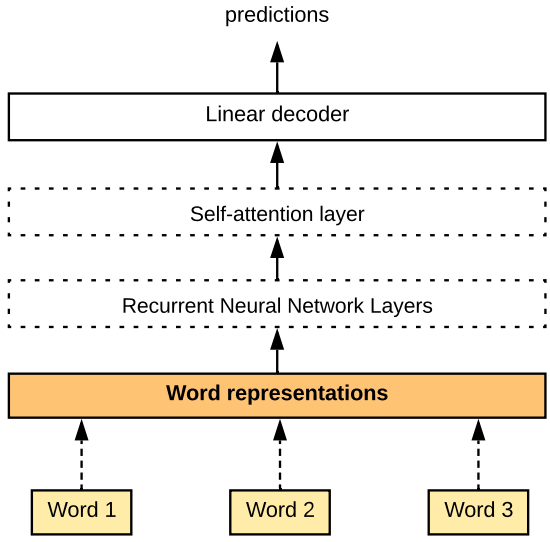

Figure 1: General neural model architecture

Word Representations

Word representations are an essential part of each neural network as embedding layer. We can consider this layer as permanent, since it is always present and we experiment only with different sizes of embedding layer and different forms of pre-trained embeddings. For this layer, we consider using standard embedding layer in the form of lookup table with dimension of 300. Different types of word representations have been recently widely used, such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). For our study, we used the pre-trained version of ELMo for Slovak language (Che et al., 2018), fastText for Slovak (Grave et al., 2018) and also pretrained word2vec for Slovak trained on prim dataset (Jazykovedný ústav Ľ. Štúra SAV, 2013).

Recurrent Neural Network Layers

We use different recurrent neural network architectures, where we consider using LSTM (Hochreiter and Schmidhuber, 1997) and Bi-LSTM (Schuster and Paliwal, 1997) with different number of stacked layers (one or two in our case). To simplify number of hyperparameters and types of architectures with different size, we consider using only size of 512.

Self-Attention Layer

To improve contribution of the most informative words, we also employ an attention mechanism. The attention mechanism assigns each word its annotation (informativeness) and the final representation is computed as weighted sum of all annotations from a sentence.

Linear Decoder

The linear decoder represents a standard linear layer, which tries to classify samples to classes. This layer can be considered as permanent, since it is always present and tries to classify samples into 2 or 3 classes depending on the target dataset.

Model Architectures

We consider several combination of described layers for evaluation of quality of neural networks for specific datasets. All architectures are shown in Table 1.

For purposes of our experiments we alternate four different word representations (randomly initialized embedding layer – LookUp, deep contextualized word representations – ELMo, fastText and word2vec). We combine different types and sizes of recurrent layers (1 LSTM, 1 Bi-LSTM) with or without use of the attention layer. For fastText and word2vec representations, we used only the last architecture employing one bidirectional LSTM with self-attention mechanism.

Model Ensemble

The last architecture we consider for improving quality of sentiment classification is using differ-

Model name	Word Representations	Recurrent Layer	Self-Attention
lookup-LSTM	LookUp	1 LSTM	None
lookup-BiLSTM	LookUp	1 Bi-LSTM	None
lookup-BiLSTM-att	LookUp	1 Bi-LSTM	Yes
ELMo-LSTM	ELMo	1 LSTM	None
ELMo-BiLSTM	ELMo	1 Bi-LSTM	None
ELMo-BiLSTM-att	ELMo	1 Bi-LSTM	Yes
w2v-BiLSTM-att	word2vec	1 Bi-LSTM	Yes
fast-BiLSTM-att	fastText	1 Bi-LSTM	Yes

Table 1: Different architectures used for experiments.

ent types of model ensemble. We consider using the same type of model for one model ensemble. Each model ensemble consists of three same models with different initialization and separate training. We also consider two types of ensemble, where models either vote for prediction or we average probabilities of model predictions.

3 Data and Evaluation

For evaluation of our models, we used two different datasets. The first dataset (*Reviews3*) consists of customer reviews of various services, which were manually labeled by 2 annotators. Since many reviews were only slightly positive or negative and agreement between annotators were not very high, we can categorize reviews into three different classes, where we consider positive, negative and neutral class (contains slightly positive or negative reviews). The second dataset (*Twitter*) consists of tweets in Slovak language (Mozetič et al., 2016), which were also labeled manually. Since some of the tweets from the original dataset did not exist anymore, we provide only evaluation on tweets available via standard Twitter API. The descriptive statistics of both datasets is shown in Table 2.

Dataset	Neg.	Neut.	Pos.	Total
Reviews3	431	2911	1978	5320
Twitter	12815	10817	27078	50710

Table 2: Statistics of used datasets.

To evaluate quality of our models we use F1 score. Since all datasets can be considered as highly unbalanced, we evaluate micro and macro F1 score separately.

One of the problems of the *Reviews3* dataset is its size. Since it contains approximately 5000 annotated reviews, we need to perform complete cross-validation, where the dataset is split in ratio 8:1:1 for train, valid and test set. For the *Twitter* dataset we split dataset in ratio 8:1:1 for train, valid and test set without any cross-validation. We also provide twitter ids for each set to preserve further reproducibility of experiments.

The only preprocessing used for our experiments is escaping punctuation to improve quality of tokenization of spaCy tokenizer in Slovak language. We also provide list of further hyper-parameters and techniques used for training our models: dropout after embedding layer 0.5; dropout after recurrent and attention layer 0.3, negative log likelihood loss, Adam optimizer.

4 Results

We performed many experiments using model architectures described in Section 2 for both datasets described in Section 3. We also compared our results with previously published results for the dataset *Reviews3* and also the dataset *Twitter*. Additionally, we also performed experiments using model ensemble for the dataset *Twitter*.

Model Results

In Table 3, we show results on the performance of the proposed models for sentiment classification for the dataset of customer reviews *Reviews3*. As we can observe, more robust models outperform smaller ones. Using deep contextualized word representations brings significant improvements of overall sentiment classification. We can also observe that a bidirectional recurrent network performs better than standard one-directional one. Using attention mechanism also brought further improvement. We also performed experiments using different pre-trained word representations with the most robust architecture. We can see that us-

ing word2vec and fastText did not bring any significant improvement for review dataset than using only randomly initialized embedding layer.

model	micro F1	macro F1
lookup-LSTM	0.7481	0.6960
lookup-BiLSTM	0.7687	0.7308
lookup-BiLSTM-att	0.7813	0.7337
ELMo-LSTM	0.8007	0.7613
ELMo-BiLSTM	0.8101	0.7681
ELMo-BiLSTM-att	0.8132	0.7693
w2v-BiLSTM-att	0.7838	0.7491
fast-BiLSTM-att	0.7819	0.7446

Table 3: Results of sentiment classification for dataset Reviews3.

In table 4, we show results on the performance of the proposed models for sentiment analysis for twitter domain (*Twitter*). We observe similar trends as for the domain of customer reviews. The most significant improvement brings using deep contextualized word representations. Similarly to the previous domain, employing bidirectional LSTM and attention mechanism improves the performance further. Unlike for dataset of customer reviews, using of fastText and word2vec representations brought improvement, which was significantly lower than using ELMo word representations.

model	micro F1	macro F1
lookup-LSTM	0.5804	0.5565
lookup-BiLSTM	0.5866	0.5614
lookup-BiLSTM-att	0.5967	0.5747
ELMo-LSTM	0.6594	0.6386
ELMo-BiLSTM	0.6671	0.6487
ELMo-BiLSTM-att	0.6978	0.6695
w2v-BiLSTM-att	0.6107	0.5908
fast-BiLSTM-att	0.6468	0.6188

Table 4: Results of sentiment classification for dataset Twitter.

Comparison with Previous Work

In Table 5, we show comparison against previously published works for sentiment classification for customer reviews. Both models used pretrained word2vec (Mikolov et al., 2013) word representations to improve quality of classification trained on prim dataset of the Slovak national corpora (Jazykovedný ústav Ľ. Štúra SAV, 2013). The first model employs SVM (Krchnavy and Simko, 2017) for sentiment classification and the second one employs neural networks along with various form of text preprocessing (Pecar et al., 2018). Since the original papers do not consider macro F1 score for evaluation, we can compare our performance only in micro F1 score. Most of our models outperforms previously published models and our best models improve overall sentiment classification by more than 6 points.

model	micro F1	macro F1
ELMo-BiLSTM-att	0.8132	0.7693
SVM baseline	0.7512	-
NN baseline	0.7296	-

Table 5: Comparison of sentiment classification for dataset Reviews 3.

In Table 6, we show comparison with the original work of the authors of dataset (Mozetič et al., 2016). The authors performed evaluation with multiple machine learning algorithms and the best one was labeled as TwoPlaneSVMbin. We cannot compare our method with theirs completely, since we were not able to obtain all samples in their dataset (due to the twitter post unavailability), hence we used only a smaller portion. We performed also experiments with another method for improving overall quality of sentiment classification – model ensemble. We trained the same model multiple times (3 in this case) and performed two types of model ensemble. In both experiments, the ensembles performed better than any of the model.

model	micro F1	macro F1
ELMo-BiLSTM-att	0.6978	0.6636
voting 3	0.6994	0.6710
mean 3	0.7008	0.6728
TwoPlaneSVMbin	0.6840*	-

Table 6: Comparison of best performing model and different types of model ensemble for dataset Twitter. * - indicates differences in used dataset

Error Analysis

In figure 2, we provide also confusion matrix of our best performed model for Twitter dataset, since our model performed much worse for the Twitter dataset than the Review3 dataset.

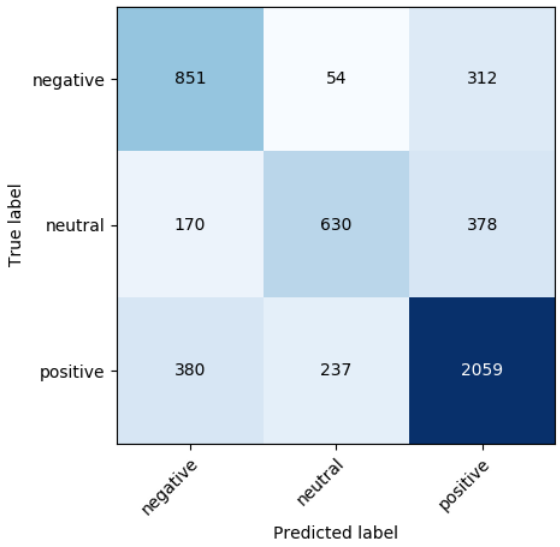

Figure 2: Confusion matrix for best performed model on Twitter dataset.

As we can observe, most mislabeled predictions are concerned with positive labels, where our model did not predict positive label or predicted it incorrectly. We performed also additional error analysis, where we looked for mislabeled tweets. After further analysis, we observed that many positively labeled tweets do not contain any sign of positive words and label was assigned due to additional information in link attached in tweet itself. This type of labeling dost not enable sentiment classification based only on textual data itself. Another observed problem could be considered labeling tweets based on real world context (e.g. political situation, twitter responses etc.), which was not provided. We suppose described problems caused significantly lower performance on Twitter dataset, since we tackled only problem of sentiment classification on texts themselves without utilizing any additional information. We believe there will be need for further manual evaluation to identify limits of human performance for this kind of dataset.

5 Conclusion

In our work, we tackled problem of sentiment classification for Slovak language, which suffers mainly from low resource datasets. We introduced several neural model architectures employing state-of-the-art techniques for sentiment analysis. As we showed, our models outperformed previously published models for sentiment classification in Slovak language. Our models performed significantly better especially for the dataset of customer reviews, where we achieved F1 score higher more than by 6 points. We suppose the main contribution to these results can be attributed to deep contextualized word representations – ELMo. Our results also showed there is only a little improvement of model performance utilizing bidirectional LSTM and attention mechanism. On the other hand, combination of those techniques along with used pre-trained word representations helps achieving significantly better results, especially for the dataset of customer reviews. The lower performance on twitter dataset could be due to nature of the dataset, where customer reviews tend to be mostly positive and negative and twitter post could be much more general in sentiment.

We suppose there is also a significant space for further improvement and application different methods, such as cross-lingual learning, where knowledge from multiple languages can be used to reduce the problem of lack of annotated resources (Pikuliak et al., 2019). Since we did not performed any significant fine-tuning and used only some of the standard setups, there can be a space to obtain even better results than we presented in this paper. Other point to consider can be training ELMo on much larger dataset, since authors of ELMo for many languages trained those representations only on the limited dataset. We provide also code for our experiments, which is available on GitHub [1].

Acknowledgments

This work was partially supported by the Slovak Research and Development Agency under the contracts No. APVV-17-0267 and No. APVV SK-IL-RD-18-0004, the Scientific Grant Agency of the Slovak Republic grants No. VG 1/0725/19 and No. VG 1/0667/18 and by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 028STU-4/2017.

References

Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sanchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing:

[1] https://github.com/SamuelPecar/Slovak-sentiment-analysis

Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Ilia Chetviorkin and Natalia Loukachevitch. 2013. Evaluating sentiment analysis systems in Russian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 12–17, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2014. Supervised sentiment analysis in czech social media. *Information Processing & Management*, 50(5):693–707.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Jazykovedný ústav Ľ. Štúra SAV. 2013. Slovenský národný korpus.

Jurgita Kapočiutė-Dzikienė, Algis Krupavičius, and Tomas Krilavičius. 2013. A comparison of approaches for sentiment classification on Lithuanian Internet comments. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 2–11, Sofia, Bulgaria. Association for Computational Linguistics.

R. Krchnavy and M. Simko. 2017. Sentiment analysis of social network posts in slovak language. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 20–25.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2018. Sentiment analysis of customer reviews: Impact of text pre-processing. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 251–256. IEEE.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. NL-FIIT at SemEval-2019 task 9: Neural model ensemble for suggestion mining. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1218–1223, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Matus Pikuliak, Marian Simko, and Maria Bielikova. 2019. Towards combining multitask and multilingual learning. In *SOFSEM 2019: Theory and Practice of Computer Science*, pages 435–446, Cham. Springer International Publishing.

Leon Rotim and Jan Šnajder. 2017. Comparison of short-text sentiment analysis methods for Croatian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 69–75, Valencia, Spain. Association for Computational Linguistics.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 24–30.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.