Sentiment Analysis of Customer Reviews: Impact of Text Pre-processing

Samuel Pecar, Marian Simko, Maria Bielikova
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Bratislava, Slovakia
{name.surname}@stuba.sk

***Abstract*—In recent years, number of customer reviews available online has grown rapidly. It is very important to give automatic support for analysis of this content. Whereas research in English language is quite covered, there is also space for research of other morphologically more complex languages like Slovak. Such languages often necessitate more advanced processing. In our work, we proposed model for sentiment analysis of customer reviews based on a neural network architecture and studied impact of several text pre-processing techniques on overall sentiment classification. We showed text pre-processing has a significant impact on performance of sentiment analysis. Though utilizing a medium-sized dataset for model training, we achieved very promising results when comparing to baseline SVM model.**

***Index Terms*—sentiment analysis, text pre-processing, neural networks**

I. Introduction

In recent years, amount of available user-generated text has rapidly grown along with ever-increasing popularity of the Web. Huge amount of text is produced by users every day, while considerable amount is focused on reviewing products or services. With increased amount information included within such text, it becomes impossible to read them (and evaluate). This is yet another vivid example of the information overload problem. For an average human, reading all available text is not possible, even if he or she read only the most relevant ones.

Automatic text summarization is viewed as a useful and important tool to help a user access information in a more efficient manner. It can help avoid necessity to read all the original documents. A specific type of text summarization is summarization of user-generated text and especially opinions within them to create so-called opinionated summaries. Many research works deal with summarization in customer reviews (e.g. product, services) [1] or comments on social networks [2]. A task of creating opinionated summaries differs from a standard summarization task due to noisy and grammatically incorrect text [3] but also due to the fact that it contains much repetitive and redundant information. Different and often opposing opinions between users are another problem: in opinion summaries, both polarities of opinions should be included. The overal polarity of summarisation should be preserved. Summaries can be very useful both for a customer but also a product owner. Decision making can be significantly supported by opinion summarization [4].

In this work, we deal with sentiment analysis in the context of opinion summarization. Sentiment analysis of customer reviews constitutes a basic step towards precise opinion summarization. Sentiment information can be considered as one of inputs for opinion summarization along with text corpora itself.

Recently, approaches based on neural networks became very popular for both summarization [5] and sentiment analysis [6]. When building a neural network model, proper inputs should be provided. In non-English languages like Slovak (and other morphologically more complex languages), initial text pre-processing is very important. However, it is much more complicated and can significantly affect performance of following steps in a processing pipeline.

In this paper, we explore the impact of pre-processing for sentiment analysis of customer reviews. We particularly focus on Slovak language. The two major contributions of our work are: (1) a neural network model for sentiment analysis for Slovak, (2) a study of impact of text pre-processing for sentiment analysis on different levels.

The rest of this paper is structured as follows. Section 2 describes state-of-the art techniques for sentiment analysis. Section 3 describes model for sentiment analysis and techniques used for pre-processing of text corpora. In section 4 we describe data and evaluation metrics used in experiments. Section 5 presents our results on sentiment analysis. Finally, section 6 discusses major observations from experiments and sets goals for our future work.

II. Related work

Sentiment analysis belongs to one of the most common tasks in the field of natural language processing. Medhat et al. in their survey discuss advantages of different methods and techniques for sentiment analysis [7]. Many researchers deal with sentiment analysis in social networks like Twitter since social networks are becoming a major source of opinionated text [8], [9]. Advances in neural networks in recent years showed feasibility of neural network models for tasks such as sentiment analysis.

Dos Santos et al. proposed a deep convolutional neural network for sentiment analysis of short texts that exploits from character to sentence level information [10]. Their convolutional neural network uses two convolutional layers to extract all the relevant features from words, which can be relevant for the task of sentiment analysis. Tang et al. introduced a neural network approach to learn continuous document representation for a sentiment detection [11]. In their approach they provide a two layer document representation, consisting of convolutional network (CNN) and long short term memory (LSTM). They produce sentence representations from the word representations. Araque et al. proposed deep learning based sentiment classifier and also combined both deep and traditional surface features [6].

Techniques employed in the pre-processing stage of text analysis such as part-of-speech tagging or dependency analysis are particularly important for sentiment analysis. Despite the importance of the mentioned techniques lying in the potential to extract relevant features for sentiment classification, Kouloumpis et al. performed experiments on Twitter sentiment analysis and claimed that part-of-speech features may not be useful for sentiment analysis in domain of microblogging [12]. In contrast with the previous work, Socher et al. proven otherwise [13]. In addition, they introduced dataset for sentiment detection, which consists of sentiment labels for phrases in the parse trees. They also introduced Recursive Neural Tensor Network, which outperforms many previous methods in several metrics on this dataset such as single sentence classification or accuracy of predicting fine-grained sentiment labels.

Information other than text can be also useful for sentiment analysis. Rosenthal et al. [14] proposed a method for sentiment analysis in Twitter using not only tweets, but also other publicly available demographic information about authors.

While research of application of neural networks in widely spread languages such as English is quiet covered, there is still insufficient research in this area for languages like Slovak [15]. One of the reasons, beside their globally limited spread and utilisation among speakers, is nature of such languages. In contrast with English, they can be more complex to process [16]. For example, Slovak language has much larger alphabet (including diacritics), richer morphology and it is inflective. It is important to research influence of such language features for sentiment analysis. To the best of our knowledge, no such study utilising neural network models has been conducted to date.

III. Model for sentiment analysis

Recurrent neural networks (RNN), especially long short term memory (LSTM) [17] can be very useful for the task of sentiment classification, especially where sentiment depends on word order in the input sentence or sentences are quite long and information about the sentiment can be forgotten in other approaches.

We propose a model consisting of LSTM cells to determine correct sentiment class 1.

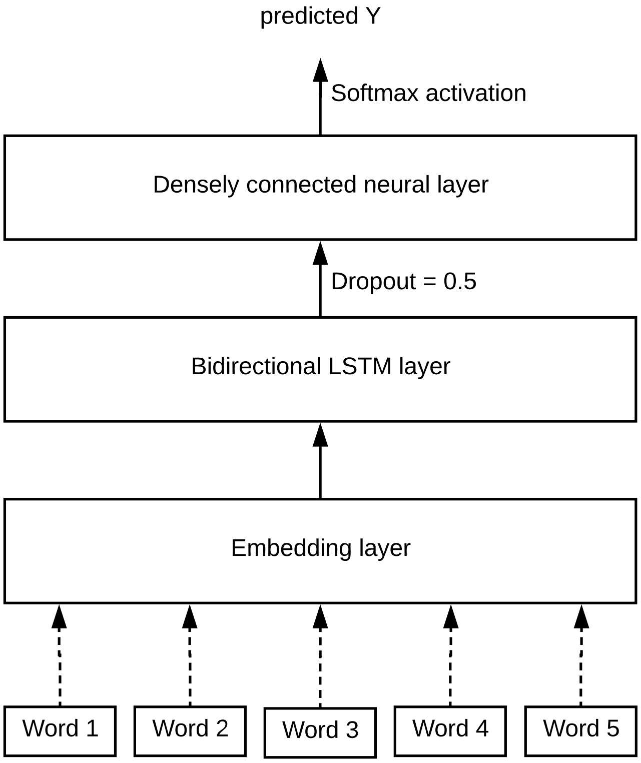

Fig. 1. Utilized neural network model

The model consists of multiple layers, where first layer codes input words via vector representation. In the next layer, we used bidirectional LSTM cells. This layer is formed by 128 units. To avoid overfitting on training dataset, we used dropout value set to 0.5. The last layer is formed by densely connected neural layer, which consists of the same number of units as is number of classes in predicted model. To predict final sentiment class the softmax activation was used.

For LSTM layer, we consider three variants:

- M1 - 1-layer LSTM
- M2 - 1-layer bidirectional LSTM
- M3 - 2-layer bidirectional LSTM

For embedding layer, we consider two possible modifications:

- E1 - dataset vocabulary with one-hot encoding
- E2 - pre-trained vectors

We describe how the input text is pre-processed for the neural network in the following section.

The model as proposed consists of layers typically used for task of sentiment analysis. At this stage of our research, our focus was on text pre-processing in order to prepare inputs for the neural network.

IV. Text Pre-processing Considerations

A very important step of sentiment analysis is pre-processing of the text. We can consider more combinations of the pre-processing parts.

The first and also the simplest one is an option using only a raw text without any pre-processing.

Another option is to tokenize text not only by splitting text with spaces but use more sophisticated tokenizer. It could correctly separate punctuation from other words. In this stage, we can consider also removing all punctuation as it has not very big impact for the task of sentiment analysis.

Many users on the Web, where the most of the opinions originate today, do not use diacritics or there can be missing diacritics in some of the words. Missing diacritics can have significant impact in the task of sentiment analysis, as some words without diacritics can have no sentiment or even opposite sentiment. A similar mistake may be caused by a typographical error, which could have a similar effect but it is much harder to correct.

The next pre-processing option is word lemmatization. Lemmatization tries to find a base form of the word and can help reduce vocabulary size. This can be quite useful for this task, but it can be argued if using lemmatized word forms along with word embedding trained on not lemmatized text is a good choice.

The last step which could be considered in the stage of pre-processing is handling emoji and emoticons. Word embeddings pre-trained on the user-generated text would be an advantage as they could introduce contextual information between emoticons and other words.

V. Data and evaluation

For evaluation of our method, we used a manually labeled dataset of reviews. The whole dataset consists of 5318 reviews of various human services which were labeled into 7 categories by two users where -3 represents most negative review and 3 the most positive ones. We took the average value of this label as a reference value. Although it is a real world dataset, the disadvantage of this dataset lies in unequal distribution of review ranking where most of the rankings are in classes 1 and 2. A complete distribution is shown in Figure 2.

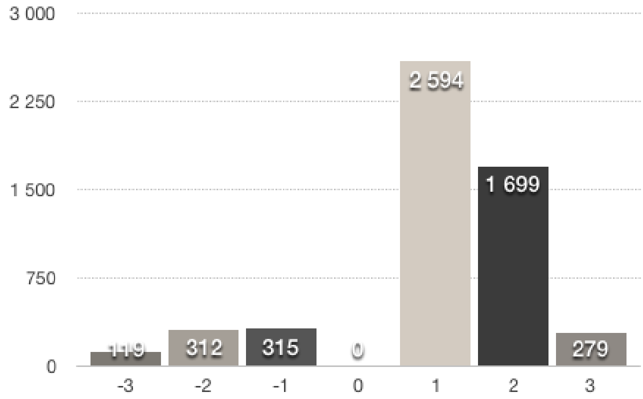

Fig. 2. Distribution of review values in dataset over 7 classes

In our experiments, we also evaluated our method by simplifying number of classes in the dataset. Classes -3, -2 and 2, 3 were merged and classes -1 and 1 were labeled as insignificant with class 0. This edited data can be viewed as a dataset for searching posts with extreme sentiment (both slightly positive and slightlly negative labels were aggregated into neutral category). A complete distribution of this altered dataset is shown in Figure 3.

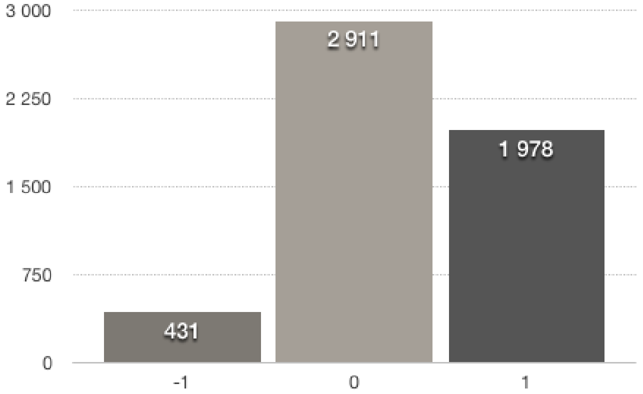

Fig. 3. Distribution of review values in dataset over 3 classes

To assess quality of our model, we focused only on the most common metric in this type of task: accuracy. Accuracy describes a ratio of correctly predicted classes and all the values in the population.

We split the available dataset into training set and test set in ratio 8:2, so 20 % of the dataset would not be used in training stage.

To evaluate our method, we used word embeddings trained on the Prim dataset created from Slovak national corpora [16], [18]. We experimented with two different dimensions for word embeddings: 80 and 300.

We compared the described neural model with other baseline model using Support Vector Machines (SVM). We show performance of our model by evaluating more setups for our neural network model and also compare model modifications with using word representation vectors with approach, where vocabulary was created only from input words.

For evaluation, we experimented with three different models as described in section III.

In experiments, we set hyperparameters as follows:

- batch size - 32,
- output embedding dimension - 80,
- optimizer - Adam,
- loss function - categorical cross entropy.

VI. Results

In order to explore impact of various pre-processing steps, we considered four different setups:

- S1 - raw text,
- S2 - tokenized text,
- S3 - tokenized text with reconstruction of diacritics,
- S4 - tokenized text with reconstruction of diacritics and removed stop words.

In Table I, we show results for sentiment analysis with these setups. To evaluate these setups, model 1 consisting of 1-layer bidirectional LSTM was used. The results showed that pre-processing has a significant impact on sentiment analysis accuracy but also showed that removal of stop words, a traditional pre-processing step from conventional methods for sentiment analysis, is not a good way to process text in task of sentiment analysis.

TABLE I
SENTIMENT CLASSIFICATION RESULTS WITH DIFFERENT SETUPS OF PRE-PROCESSING

	classes	S1	S2	S3	S4
E1	3	0.5733	0.6082	0.6824	0.5989
	7	0.4830	0.5218	0.6607	0.6447
E2	3	0.6127	0.6512	0.7296	0.6250
	7	0.6193	0.6784	0.7156	0.6682

We took the best performing setup (S3) and were further interested how other model variants (M1-M3) will perform. In Table II, we show results for sentiment classification into 3 and 7 classes with different models and also embedding layer.

TABLE II
SENTIMENT CLASSIFICATION RESULTS UTILIZING MODELS

	classes	M1	M2	M3
E1	3	0.4821	0.6824	0.6539
	7	0.4521	0.6607	0.6447
E2	3	0.5338	0.7296	0.6903
	7	0.4906	0.7156	0.6835

In Table III, we show our results for sentiment classification into 3 and 7 classes in comparison to models trained using Support Vector Machines (SVM).

TABLE III
COMPARISON AGAINST BASELINE SVM MODELS

	E1		E2	
classes	SVM	M2	SVM	M2
3	0.7347	0.6824	0.7512	0.7296
7	0.7124	0.6607	0.6947	0.7156

The results obtained from our experiments show that employing word embeddings improved overall accuracy in all models. We can also see, that one directional LSTM reached much worse results than the bidirectional ones. In our experiments, 1-layer bidirectional LSTM obtained better results than 2-layer. This observation could be caused due to small number of epochs, but also quite small dataset. Our best model showed promising results and in comparison with SVM model obtained the best results in sentiment classification to 7 classes.

We see the further potential to improve results in line with improving the neural network training phase. Currently, due to the rather limited dataset size, training accuracy over 20 epochs stopped at accuracy $\tilde{0}.83$. We expect that with more data the neural network would be able to learn more.

VII. CONCLUSION AND FUTURE WORK

In our research work, we tackle sentiment analysis in the context of opinion summarization. Our long term goal is to devise methods for opininon summarisation in user created text corpora like customer reviews. We expect from such methods to correctly deal with information of different polarity, which would be neglected if traditional text summarisation methods were simply employed. Hence, we consider sentiment analysis as a basic and crucial step of summarization.

In this paper, we were particularly focused on text pre-processing step, which is very important for morphologically rich inflective languages like Slovak. We proposed a LSTM neural network model and explored various pre-processing variants providing inputs for the model which would result into the most accurate sentiment classification.

Our experiments showed that LSTM neural network is useful for sentiment classification in Slovak language despite its complexity. Results of our experiments showed very promising results, which can be a starting point for future work.

In the future work, we would like to further explore different setups of pre-processing and their impact on sentiment classification. The reported experiments involved only 20 epochs of training, yet they showed promising results. We expect even better results if we would increase number of epochs significantly. More data are however necessary. This represents a non-trivial obstacle when working with minor language like Slovak. Our effort will be focused primarily on this task. Our aim is assembly of a bigger dataset of Slovak user-generated texts with more balanced distribution of sentiment classes, which has a huge impact on final accuracy especially in classification of negative sentiment, where total volume was the smallest.

We also plan to elaborate more on lemmatization and emoticon processing as we outlined in section IV. Another point to discuss is usage of word embeddings. To obtain better results on sentiment analysis we would need to get word vectors trained on similar texts as customer reviews containing both emojis and emoticons.

REFERENCES

[1] Y.-H. Hu, Y.-L. Chen, and H.-L. Chou, “Opinion mining from online hotel reviews – A text summarization approach,” *Information Processing & Management*, vol. 53, no. 2, pp. 436–449, mar 2017.

[2] L. Wang, H. Raghavan, C. Cardie, and V. Castelli, “Query-focused opinion summarization for user-generated content,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1660–1669.

[3] G. Murray, E. Hoque, and G. Carenini, “Chapter 11 - opinion summarization and visualization,” in *Sentiment Analysis in Social Networks*, F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, Eds. Morgan Kaufmann, 2017, pp. 171 – 187.

[4] X. Yuan, N. Sa, G. Begany, and H. Yang, “What users prefer and why: A user study on effective presentation styles of opinion summarization,” in *Human-Computer Interaction – INTERACT 2015*. Springer International Publishing, 2015, pp. 249–264.

[5] A. M. Rush, S. Chopra, and J. Weston, “A Neural Attention Model for Abstractive Sentence Summarization,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, sep 2015, pp. 379–389.

[6] O. Araque, I. Corcuera-Platas, J. F. S°nchez-Rada, and C. A. Iglesias, “Enhancing deep learning sentiment analysis with ensemble techniques in social applications,” *Expert Systems with Applications*, vol. 77, pp. 236 – 246, 2017.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[8] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010.

[9] P. Korenek and M. Simko, "Sentiment analysis on microblog utilizing appraisal theory," *World Wide Web*, vol. 17, no. 4, pp. 847–867, 2014.

[10] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014, pp. 69–78.

[11] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1422–1432.

[12] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!" in *ICWSM*. The AAAI Press, 2011.

[13] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, October 2013, pp. 1631–1642.

[14] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, 2017, pp. 502–518.

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

[16] L. Gallay and M. Simko, "Utilizing vector models for automatic text lemmatization," in *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 2016, pp. 532–543.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. ústav Ľ. Štúra SAV, "Slovenský národný korpus – prim-6.1-public-sane," http://korpus.juls.savba.sk, 2013.