Sentiment Analysis for Multilingual Corpora

Svitlana Galeshchuk
Governance Analytics,
PSL Research University /
University Paris Dauphine
Place du Marechal
de Lattre de Tassigny,
75016 Paris, France
s.galeshchuk@gmail.com

Julien Jourdan
PSL Research University /
University Paris Dauphine
Place du Marechal
de Lattre de Tassigny,
75016 Paris, France
julien.jourdan@dauphine.psl.eu

Ju Qiu
Governance Analytics,
PSL Research University /
University Paris Dauphine
Place du Marechal
de Lattre de Tassigny,
75016 Paris, France
ju.qiu@dauphine.psl.eu

Abstract

The paper presents a generic approach to the supervised sentiment analysis of social media content in foreign languages. The method proposes translating documents from the original language to English with Google's Neural Translation Model. The resulted texts are then converted to vectors by averaging the vectorial representation of words derived from a pre-trained Word2Vec English model. Testing the approach with several machine learning methods on Polish, Slovenian and Croatian Twitter corpora returns up to 86 % of classification accuracy on the out-of-sample data.

1 Introduction

Sentiment analysis is gaining prominence as a topic of research in different areas of application (journalism, political science, marketing, finance, etc.). In the last two decades, opinion-rich data sources are widely available because of web-resources and social networks. While lexicon-based frameworks have long been investigated for sentiment analysis, deep learning methods with a vectorial representation of words are proving to deliver promising results. The integration of two types of methods is widely investigated as well. Thus, the sentiment analysis approaches usually require either fine-grained lexicon of most frequent words along with their polarity scores or the dataset large enough for supervised training of deep learning network, sufficient computational memory, etc.

Moreover, most of the open-source datasets for training sentiment models comprise English-language texts. The lexicons are not always available for other languages, and it remains a time-consuming task to construct them. It motivates us to build on the existing approaches and test a rather general method to run a sentiment analysis for different languages without polarity dictionaries using relatively small datasets.

We address the challenges for sentiment analysis in Slavic languages by using the averaged vectors for each word in a document translated in English. The vectors derive from the Word2Vec model pre-trained on Google news.

Researchers tend to use the language-specific pre-trained Word2Vec models (e.g., Word2Vec model pre-trained on Wikipedia corpus in Greek, Swedish, etc.). On the contrary, *we propose benefiting from Google's Neural Translation Model translating the texts from other languages to English. Translated documents are then converted to the fixed-vectorial representation with Google Word2Vec model.*[1] *The supervised machine learning classifiers such as Gradient Boosting Trees, Random Forest, Support Vector Machines provide sufficiently high accuracy on the out-of-sample data converted to the aggregate vectors.*

The rest of the paper is structured as follows: Section 2 provides a brief review of related literature. Section 3 describes the methodology. Section 4 expands on data used. Section 5 presents the results from our experiments. Section 6 concludes with some observations on our findings and identifies directions of future research.

2 Related Work

This section elaborates on existing methods for sentiments analysis and the adjacent approaches to text data treatment that have helped us formulate the proposed process of sentiment analysis. Following Dashtipour et al. (2016), we divide sentiment analysis systems on lexicon-based, corpus-based and hybrid.

[1] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 120–125,
Florence, Italy, 2 August 2019.

2.1 Lexicon-Based Methods

Lexicon-based methods employ the dictionaries of pre-defined words with corresponding polarity scores. These scores define how positive the term is. Some approaches (e.g., Vader lexicon in Hutto and Gilbert (2014)) use the opinion of several experts and the final polarity measure equals the mean of the corresponding scores. A subset of the most popular and promising lexicon-based sentiment classifiers for English corpora has been reported in Levallois (2013). Concerning Slavic languages, Slovak lexicon translated from English and annotated with Bare-bones particle swarm optimization helps achieve 0.865 F1 score in sentiment classification reported in Krchnavy and Simko (2017). Gombar et al. (2017) construct Croatian domain-specific lexicon for domain-specific classification; Haniewicz et al. (2013) run sentiment analysis with polarity lexicon for reviews in Polish that renders up to 79% of accuracy. We will refer to these papers later in our study to corroborate our results by comparison with the existing methods.

The idea proposed in Wan (2009) shares some similarities with our method. Authors translate Chinese text in English and then employ lexicon-based ensemble method to classify texts on positive or negative. The reported accuracy is 85.3% though it requires Chinese and English lexicon and some additional calculation to create the ensemble method. However, word scoring in each constructed lexicon usually relies on human treatment and perception. The task is also labor-intensive, and it may be challenging to find fine-grained lexicons for some languages.

Moreover, a well-known drawback of lexicon-based method is the contextual ignorance as some terms may have different meanings in various documents. Besides, some documents (e.g., short texts as tweets) sometimes do not include any word from the lexicon. The introduction of word vectorial representation tend to address this disadvantage.

2.2 Corpus-based and Hybrid Methods with Vectorial Representation

Embedding approaches usually rely on the semantic vector spaces derived from the neural networks. Their application in supervised experimental setups for polarity analysis often demonstrates superior performance to the lexicon-based methods (Le and Mikolov, 2014; Severyn and Moschitti, 2015). As the reference point in our study we use the papers of Giatsoglou et al. (2017) and Garten et al. (2018) where authors meticulously employ Sentence2Vec in their methodological settings. Giatsoglou et al. (2017) uses Sentence2Vec based on the Word2Vec model learned from the Wikipedia corpus in Greek. The performance is evaluated with the datasets of mobile phonesâĂŹ reviews in Greek. The model that exploits the vectors derived from the Wikipedia corpora+the reviews provides the highest accuracy of 70.89-82.40% on test samples. Author further try hybrid methods (lexicon- and embedding- driven) that deliver slightly better results. Rotim and Šnajder (2017) use similar approach for Croatian corpora obtaining 0.822 as F1 score for game reviews but the results are much worse for the Twitter dataset. In contrast to the authors, we do not train our Word2Vec model for the corpora in Slavic languages. Instead, we employ the pre-trained Google News Word2Vec model after translating texts to English. It makes our approach more universal and easier to apply to the foreign language corpora yielding satisfactory accuracy (see Results).

Garten et al. (2018) compute the cosine similarity between the aggregate vectorial representation of documents and the âĂIJnegativeâĂİ and âĂIJpositiveâĂİ dictionaries. Precision on the IMBD English reviews data varies between 0.70-0.75. Our findings show that the introduction of the polarity dictionaries delivers less accurate outputs than using Sentence2Vec. However, our set-up does not foresee unsupervised learning.

Feature-based approach for Czech language sentiment classification renders 0.69 as F1 measure in Habernal et al. (2013). The method has to be adjusted for other languages if used.

Zhang et al. (2017) report another approach for Twitter sentiment classification employing character-based Convolutional neural networks with different languages. The method transforms the characters in alphabetic order in UTF-8 codes facilitating sentence segmentation. The character embedding matrix is then used as an input for the convolutional neural network. We consider these findings as one of the benchmarks for comparison in our study.

3 Methodology

Recall from the previous sections that we tend to develop a sentiment analysis approach for multi-language use. Fig.1 depicts the proposed method.[2]

```
1.For each document in available collection:
  Initialize text preprocessing[2]
  Translate to English
2. Load Google Word2Vec model
3. For each document in translated texts:
    For each word in document:
     If word is in Google Word2Vec vocabulary:
      Draw vector representation
    Compute aggregated representation of words in document
4. Generate the file with aggregated vector representation for each document.
5. Initialize machine learning method.
```

Figure 1: Generic process of the multilingual sentiment classiffication

Word vectorial representation spurs the interest of many researchers in natural language processing due to its capacity to catch the meaning of terms depending on the context. Researchers and practitioners who use deep learning methods for sentiment analysis tend to learn the embedding from the available dataset or pre-trained word embedding models (i.e., Word2Vec, Glove). Each document is then represented as the stack of vectors. Document padding unifies the number of vectors that serve as the network input. It means that the network deals with [batchsize*size of vectorial stack*number of features] input data dimensions which may be computationally costly.

Instead, our model exploits vectorial representation of the words with a transfer learning approach: Google Word2Vec pre-trained model serves as a source of 300-dimensional dense vectors for each word in the text. Then the model computes an elementwise sum of the vectors divided by the number of terms in the text.

The use of Google Word2Vec model has several advantages over learning embeddings from the training data: (i) Google model has been pre-trained with the corpora of the news containing circa 100 billion words where each term has been used more than once and in different contexts. It makes the model the state-of-the-art regarding the quality of vectors which plays a crucial role in our study as we use the translated text from Slavic languages to English; (ii) Google model comprises approximately 3 million words and phrases. This vocabulary covers the lion share of lexicon employed by web-resources and social networks users; (iii) we do not need to construct a large dataset to train our model as the vectors have already been pre-trained with a significant number of terms.

Google Translation. Machine translation does not always provide perfect accuracy from the linguistic point of view. However, the resulted translation with recently introduced GoogleâĂŹs Neural Machine Translation approach tends to deliver English text contextually similar to the input document (see Wu et al. (2016) for more details).

3.1 Machine Learning Methods Used

This subsection discusses the machine learning classifiers employed in a supervised learning approach to classify texts on positive or negative. The implementation details are stored in the Github repository.

Support Vector Machines Classification. This approach belongs to the family of versatile machine learning methods with high accuracy on non-large datasets. It tries to find the broadest possible margin between positive and negative classes. As in Giatsoglou et al. (2017) we use linear Support Vector Classifier (SVC) and Gaussian Radial Basis Function (RBF) in our set-up.

Random forest (RF) helps overcome the disadvantages of a single decision tree by summarizing and averaging predictions over the number of trees. It is an ensemble learning approach that uses the outputs of the individual predictors as votes. If the positive class gets more votes, the method will return the corresponding result.

Gradient Boosting Trees. In our set-up Gradient boosting (GBT) method represents an ensemble of classification decision trees. Each tree sequentially joins the ensemble correcting the antecedent by fitting its residual errors.

Deep Neural Networks (DNN) are versatile methods that address complex machine learning tasks. They are effective to capture non-linearities in the data and latent relationships between the variables. We build on state-of-art DNN architectures and recent findings on hyperparameters cali-

[2]We removed urls, emojis, digits and punctuation marks as text preprocessing

bration in our empirical search for the model with the best possible accuracy. The architecture of our DNN comprises 2 hidden layers with dropout rate of 0.2

3.2 Evaluation

For this classification problem, the quality of a model is measured by the proportion of correctly classified observations (accuracy). The receiver operating characteristic curve plots true positive rate against false positive rate for the test set. The area under the curve (AUC) represents another way to compare the classifiers.

4 Data

Multilanguage Corpora. We use the corpora in Polish, Slovenian and Croatian in our experimental set-up. Polish, Slovenian and Croatian languages belong to the Indo-European family as well as English. However, they are members of the Slavic branch that makes these languages share less close ties with English than, for example, French, Spanish or Italian would have.

We retrieve the dataset with texts and corresponding polarity scores for tweets in mentioned languages from the website of the European research infrastructure CLARIN. [3] Our dataset comprises 2794 tweets in Polish (1397 positive and 1397 negative), 4272 tweets in Slovenian (2312 positive and 1950 negative) and 3554 tweets in Croatian (2129 positive and 1425 negative).

Tweets may contain emojis and/or URLs. We removed them together with digits and punctuation marks as a part of the data preprocessing step. Later we employ Google Translator API via Python Library Google-api-translate to translate texts to English language. Google Translation Library is easy to use and it takes approximately 12 min to translate 1000 tweets. The corpora for Slovenian and Croatian languages are imbalanced. Hence, we use stratified split in the cross-validation settings that returns folds with the corresponding ratio of classes in training and test sets.

5 Results

After mentioned datasets have been translated to English, each tweet is converted to the vector creating the averaged representation of the document words. Data is split into training/testing sets as 80/20 respecting the ratio of positive and negative observations.

Table 1 presents the accuracy results for three datasets (Polish, Slovenian, Croatian). The overall accuracy of the best classifiers is more than 76% which may be seen as satisfactory taking into consideration state-of-the-art findings. The tree-based methods usually deliver marginally better accuracy to the SVC and DNN classifiers. The accuracy of more than 78% is higher than the one reported in Garten et al. (2018) where authors use both documents aggregate representation and predefined lexicon. Recall measure is lower for the negative class in case of the corpora in Croatian and Slovenian. The issue of the imbalanced data may explain it.

Data/cl.	RF	GBT	SVM	RBF	DNN
Polish	76.30	**78.46**	76.84	73.25	75.58
Slovenian	73.12	**76.10**	75.70	71.80	75.69
Croatian	**86.32**	86.26	85.17	86.14	85.58

Table 1: Accuracy with applied machine learning methods

Fig. 2, 3, 4. depict the ROC curve with the largest AUC for each datasets (2) RF for Polish, (3) GBR for Slovenian, (4) RF for Croatian. ROC curve detects tree-base methods outperform the rest of approaches for all datasets.

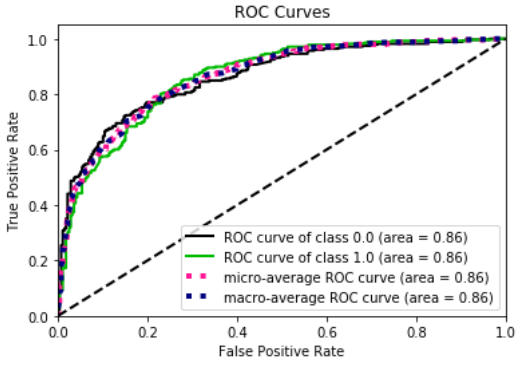

Figure 2: ROC curve with the largest AUC for Polish language

5.1 Comparison of Methods

In the previous sections we have described part of the exisiting state-of-the-art methods. This subsection tries assessing the results obtained with our approach by comparison with mentioned studies. However, the direct juxtaposition is restricted as authors use different languages, corpora and

[3] https://www.clarin.si/repository/xmlui/

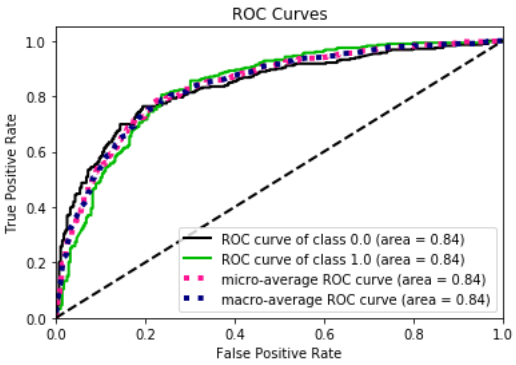

Figure 3: ROC curve with the largest AUC for Slovenian language

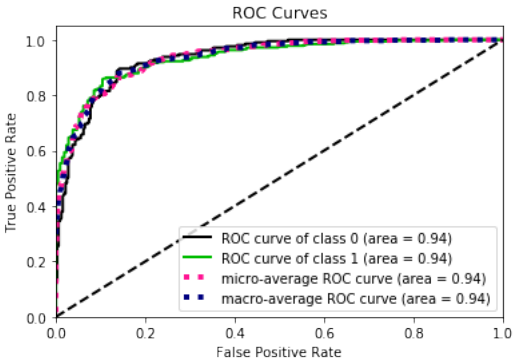

Figure 4: ROC curve with the largest AUC for Croatian language

scores to evaluate classifiers. Thus we choose the papers with developed methods in Polish, Slovenian and Croatian tested on social media or product review corpora to make the quantitative comparison as unbiased as possible. Table 2 reports the evaluation findings. The analysis proves that despite being generic our approach returns similar or better results for sentiment analysis in Polish, Slovenian and Croatian comparing to the other methods.

Author	Measure	Reported results	Results with developed method
POLISH LANGUAGE:			
Haniewicz et al. (2013)	Accuracy	circa 79.00	78.46
Zhang et al. (2017)	Accuracy	81.19	78.46
Buczynski and Wawer (2008)	Accuracy	77.05	78.46
SLOVENIAN LANGUAGE:			
Zhang et al. (2017)	Accuracy	78.07	76.10
Kadunc (2016)	Accuracy	76,20	76.10
CROATIAN LANGUAGE:			
Gombar et al. (2017)	F1 Score	0.66	0.86
Rotim and Šnajder (2017)	F1 Score	0.57	0.86
Agić et al. (2010)	F1 Score	0.63	0.86

Table 2: Comparison of Findings

6 Conclusion

The paper introduces and elaborates on the developed generic approach to sentiment analysis of multilingual corpora that encompasses translating texts to English, aggregating vectorial representation of translated words and eventually applying machine learning methods to classify documents on positive or negative. As pointed out earlier, the aim of the study is not to compete with the existing techics in terms of accuracy but to propose method that does not suffer from one-language applicability and is simple to implement. We build on the state-of-the-art and present a general set-up which may be used in supervised sentiment analysis for different Slavic languages. Testing the accuracy of our approach on a collection of tweets in three Slavic languages delivers comparable accuracy to the reported findings from recent papers on sentiment analysis for English and non-English corpora (see Related Work and Comparison of Methods). However, the difference in the classification accuracy for Polish, Slovenian and Croatian languages motivates us to test the method with other Slavic languages. These discrepancies may arrise from the quality of translation as well as from the imperfections in labeling the data. We are working on our own pre-labeled balanced dataset to further improve the approach.

Acknowledgments

We are grateful for the received support from the research initiative "Governance Analytics" funded by the PSL University under the program "Investissements Avenir" launched by the French Government and implemented by ANR with the references ANR-10-IDEX-0001-02 PSL Paris Sciences et Lettres (PSL). We would also like to thank the anonymous reviewers for their suggestions and comments.

References

Željko Agić, Nikola Ljubešić, and Marko Tadić. 2010. Towards sentiment analysis of financial texts in croatian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Aleksander Buczynski and Aleksander Wawer. 2008. Shallow parsing in sentiment analysis of product reviews. In *Proceedings of the Partial Parsing workshop at LREC*, volume 2008, pages 14–18.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Justin Garten, Joe Hoover, Kate M Johnson, Reihane Boghrati, Carol Iskiwitch, and Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, 50(1):344–361.

Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch Chatzisavvas. 2017. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224.

Paula Gombar, Zoran Medić, Domagoj Alagić, and Jan Šnajder. 2017. Debunking sentiment lexicons: A case of domain-specific sentiment classification for croatian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 54–59.

Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 2013. Sentiment analysis in czech social media using supervised machine learning. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 65–74.

Konstanty Haniewicz, Wojciech Rutkowski, Magdalena Adamczyk, and Monika Kaczmarek. 2013. Towards the lexicon-based sentiment analysis of polish texts: Polarity lexicon. In *International Conference on Computational Collective Intelligence*, pages 286–295. Springer.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Klemen Kadunc. 2016. Using machine learning for sentiment analysis of slovene web commentaries. *University of Ljubljana*.

Rastislav Krchnavy and Marian 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. IEEE.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Clement Levallois. 2013. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 414–417.

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.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Shiwei Zhang, Xiuzhen Zhang, and Jeffrey Chan. 2017. A word-character convolutional neural network for language-agnostic twitter sentiment analysis. In *Proceedings of the 22nd Australasian Document Computing Symposium*, page 12. ACM.