Aspect-based sentiment analysis

Jaklič, Žan
63130073
zj8850@student-uni.lj.si

Ramovš, Iztok
63130204
ir8617@student-uni.lj.si

Mav, Matjaž
63130148
mm3058@student-uni.lj.si

1 Introduction

A lot of sentiment analysis has been done on short texts, e.g. on tweets, yelp reviews, amazon reviews, and some of the research has been done on longer texts like news, blog posts, etc. There is a lot of incentive to use sentiment analysis on those cases. Such techniques can for example help companies and researchers with understanding of user opinions or filter out unimportant ones. One example for sentiment analysis application can even be predicting how markets will shift due to financial news.

On one side there is interest to assign sentiment to the whole document and on the other, there is interest to assign sentiment to each entity mentioned in the document separately. We will focus on the more scarcely researched second option, called aspect based sentiment analysis, processing a dataset constructed of Slovene news article.

2 Related work

We reviewed the proposed literature and some additional papers, which research the problem of aspect-based sentiment analysis. Across all papers, the majority of recently popular approaches for natural language processing were used.

First, we read a paper (Asghar et al., 2014), which mostly focuses on data preprocessing and feature extraction. This paper along with most of the others we read recommends heavy preprocessing of the initial texts, including punctuation and stop word removal, transforming letters to lowercase, part of speech tagging and lemmatization. However (Bučar, 2017), which analyzes Slovene documents, recommends that we omit stop word and capitalization removal.

Most of the approaches take advantage of machine learning models, while (Sweeney and Padmanabhan, 2017) also tried predicting sentiment with lexicons. (Biyani et al., 2015) divided the classification problem into several binary classifications, where (Ding et al., 2018) used multiclass classification approach. Machine learning models used for classification ranged from Naïve Bayes (Bučar, 2017), Support Vector Machines (Tang et al., 2016), to neural networks (Jebbara, 2016) and transformers (Yang et al., 2019). There were also a few novel approaches. (Jebbara, 2016) created custom sentiment embeddings, (Wallaart and Frasincar, 2019) took the ontology approach, where they transformed words into aspects and classified sentiment based on domain specific rules. (Guha et al., 2015) used a special form of clustering instead of word embeddings. (Hercig et al., 2016) approached the problem with unsupervised methods. Intel's NLP architect, which is based on (Mamou et al., 2018), uses semi-supervised learning, where a domain specific opinion lexicon is automatically created and can be then manually corrected if needed.

Besides (Bučar, 2017) we also reviewed another lexicon approach on Slovene texts (Kadunc and Robnik-Šikonja, 2017). We have not had the time to go through the doctoral dissertation (Bučar, 2017), but we intend to use it for further guidance in our research project.

3 Data description

The corpus we will use is the SentiCoref 1.0 (Žitnik, 2019) which consists of 837 documents and 433 thousand tokens which were selected from SentiNews 1.0 corpus (Bučar, 2017). The text contents of those selected documents are from five different Slovene news portals. The SentiCoref corpus is already annotated with following data for each token:

- location of token in text,
- type of named entities (person, organization or location),

- coreference to named entities,
- discrete sentiment range from very negative to very positive for each entity.

Most of the annotated sentiment in database is tagged as neutral; the distribution for entities is following: 30 Very negative, 1801 Negative, 10869 Neutral, 1705 Positive and 24 Very positive.

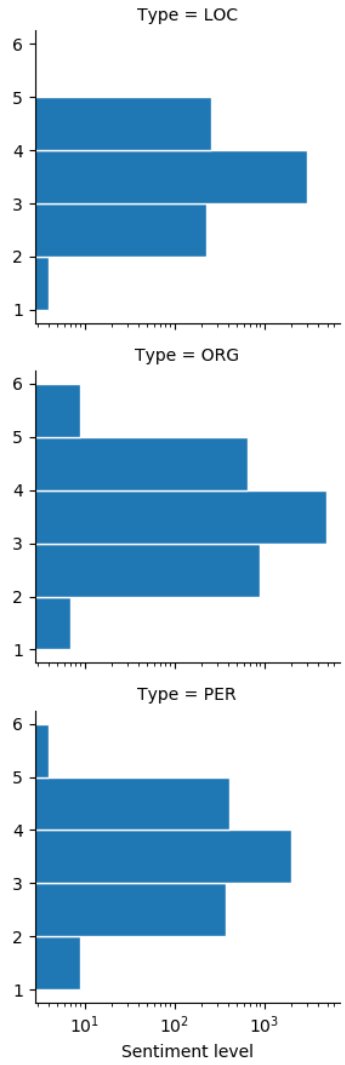

Figure 1: Sentiment distribution compared to the entity type. Note that frequencies are in a logarithmic scale.

4 Project idea

At the moment we have limited knowledge regarding machine learning models used for advanced language processing. We have therefore decided that we will come to the best options regarding preprocessing, feature extraction, modelling and evaluation through experimental testing. We intend to implement most of the perspective methods studied in related works and will try to develop a few of our own solutions in each of the steps mentioned above.

Our goal is to first build a simple model, so we get familiar with the data and algorithms used, and then improve every process in the pipeline to achieve best results possible.

We will initially use less preprocessed text, as that appeared to yield better results in Slovene texts. We aim to extract the local context of an entity with taking into account a fixed amount of tokens in its neighborhood. Bigrams and word embeddings appeared to be useful for evaluating sentiment in previous research, so we will use them as well. We will use Slovene lexicons to establish a benchmark, and compare them to simpler models, such as Naive Bayes, first. Most of the previous papers used accuracy and F-score for evaluating their results. However we are concerned about the distribution of our sentiment classes, where very polar classes represent a very small number of all cases.

We aim to improve our results throughout development, with experiments, as we gain more knowledge about our dataset and language processing procedures.

4.1 Preprocessing

- Parsing initial data, tokenization
- Punctuation and capitalization processing
- Part of speech tagging
- Lemmatization
- Stop word removal

4.2 Data inspection

- Class distribution among entities
- Source and domain bias

4.3 Feature extraction

- Local context
- Negation, conjuction, superlative evaluation
- Punctuation
- Ontology based approach
- N-grams
- TF-IDF, word embeddings
- Additional data to expand feature space

4.4 Modelling and evaluation

- Lexicons
- Simpler machine learning models
- Deep learning models
- Transformer models
- Stacking
- Accuracy, F-score

4.5 Main question for future work

- What is the proper amount of text preprocessing?
- How should we evaluate local context of an entity?
- Is our dataset large enough for deep learning and transformer models?
- How important are the very polar sentiment classes, which are very small?

References

Dr. Muhammad Asghar, Aurangzeb Khan, Shakeel Ahmad, and Fazal Kundi. 2014. A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Research International*, 4:181–186.

Prakhar Biyani, Cornelia Caragea, and Narayan Bhamidipati. 2015. Entity-specific sentiment classification of yahoo news comments.

Jože Bučar. 2017. Manually sentiment annotated slovenian news corpus SentiNews 1.0. Slovenian language resource repository CLARIN.SI.

Jože Bučar. 2017. *Sentiment based classification of the web texts*. PhD dissertation, Fakulteta za informacijske študije v Novem mestu.

Jin Ding, Hailong Sun, Xu Wang, and Xudong Liu. 2018. Entity-level sentiment analysis of issue comments. pages 7–13.

Satarupa Guha, Aditya Joshi, and Vasudeva Varma. 2015. SIEL: Aspect based sentiment analysis in reviews. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 759–766, Denver, Colorado. Association for Computational Linguistics.

Tomáš Hercig, Tomáš Brychcín, Lukáš Svoboda, Michal Konkol, and Josef Steinberger. 2016. Unsupervised methods to improve aspect-based sentiment analysis in czech. *Computacion y Sistemas*, 20:365–375.

Soufian Jebbara. 2016. Aspect-based sentiment analysis using a two-step neural network architecture. volume 641, pages 153–167.

Klemen Kadunc and Marko Robnik-Šikonja. 2017. Slovene sentiment lexicon KSS 1.1. Slovenian language resource repository CLARIN.SI.

Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat. 2018. Term set expansion based NLP architect by intel AI lab. *CoRR*, abs/1808.08953.

Colm Sweeney and Deepak Padmanabhan. 2017. Multi-entity sentiment analysis using entity-level feature extraction and word embeddings approach. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 733–740, Varna, Bulgaria. INCOMA Ltd.

D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):496–509.

Olaf Wallaart and Flavius Frasincar. 2019. *A Hybrid Approach for Aspect-Based Sentiment Analysis Using a Lexicalized Domain Ontology and Attentional Neural Models*.

Heng Yang, Biqing Zeng, JianHao Yang, Youwei Song, and Ruyang Xu. 2019. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction.

Slavko Žitnik. 2019. Slovene corpus for aspect-based sentiment analysis - SentiCoref 1.0. Slovenian language resource repository CLARIN.SI.