**Sentiment Frames Extraction**

A Research Proposal Presented to

The Humanities Department

by

**Nikita Y. Shchennikov**

National Research University Higher School of Economics

School of Linguistics

Academic advisor:

Assistant Professor S.Y. Toldova

Moscow 2020

## Abstract

Sentiment analysis is the highly demanded and rapidly developing branch of the Natural Language Processing. It is widely used in different fields of the NLP. Unfortunately, there are considerable problems in some studies in this branch. Most of them are very narrowly specialized and produce results that are not applicable to a variety of tasks. We will broaden the existing golden standard sentiment lexicon using sentiment frames. Тhen based on received information makes sentiment analysis of non-topical short texts that were collected from the Russian segment of Twitter. Our purpose is to make a model for this analysis based on different approaches that could be used on non-topical texts with rather high accuracy. Due to the fact that most of the previous works are based on thematic datasets, this paper should shed some light on this problem with a different to a standard approach. Despite increasing popularity and usability of such Transformers as BERT and Open AI GPT-2 and their outstanding text predictions, sentiment analysis of non-topical texts is still rather a hard task even for them. We believe that this new approach could give us information that could be useful in solving this problem. Usage of sentiment frames combined with modern solutions for classification and embeddings has the potential to be a fresh look at this problem.

## Introduction

Sentiment analysis is the task of automatic extraction of text sentiment. The main idea of sentiment analysis is to assign a certain value to each text, that will represent its “polarity”.

The assignment method could vary from simple binary “0” for the negative sentiment and “1” for a positive [14] to the scale where “1” is a lower boundary and represents a highly negative sentiment and “5” is an upper boundary [15], which represents a highly positive sentiment.

Nowadays, sentiment analysis is one of the leading tasks in Natural Language Processing. There are many pieces of research from big companies in this field [21][16], because sentiment analysis is increasingly becoming an important factor in the work of politics, news and reviews aggregators, banking systems. The main problem in sentiment analysis that it strongly relies on thematic datasets such as film or book reviews, topical news, product reviews, which make most of the models narrowly targeted. That imposes certain constraints on future researches. Data that was collected in the sentiment analysis of film reviews could be heavily used in the analysis of political news. Lexicons could become a problem in sentiment analysis as well. Usually, they are nonrepresentational or thematic, and that leads us to the same problem which was described earlier.

In this research, we will implement the method of sentiment framework in a sentiment analysis, which is based on the work of Karnaukhova, Loukachevitch [13]. Sentiment framework is the verb-mediated model of specific connections of predicates that relies on the idea that the predicate could affect the polarity of the subject and object of the sentence (e.g. “X wins Y” make X positive and Y negative). According to Deng and Wiebe [12], the verb could assign certain sentiment on the words that are connected to it. So, the frame acts as a set of encyclopedic, linguistic and cognitive knowledge.

Here, we choose the corpus of short texts in Russian based on Twitter [3] as our train data. Using this dataset we are solving two main problems: objectivity of the text and outlining the subject of the statement. The first problem is raised in the work [2] of Bo Pang and Lillian Lee. They made an assumption that the emotionality and subjectivity of the text make polarity predictions more accurate. After they split their data on subjective and objective, they analyzed the subjective and obtained more precise results. The second problem dates back to the method of sentiment frames. Using tweets as data will grant exemption from defining the subject of the statement since it will be an author of the text on most of the occasions. So we put forward a hypothesis that tweets combined with sentiment framework should be a good base for polarity predictions.

Taking all the above into consideration we will try to implement this idea into the text preprocessing and improve automatic sentiment analysis models of non-specific language basing on the corpus of Russian texts from Twitter.

## Literature Review

Sentiment analysis is the task of opinion mining from text and extracting subjective data. It helps a business to understand the social attitude towards their product. Since the very first works in this field [4], it took a long way to develop brand new and highly productive deep learning algorithms like BERT (Bidirectional Encoder Representations from Transformers) [7] or CSS (Contextual Semantic Search) [22]. In the beginning, it was just an approach in computer content analysis which was quantifying patterns in text and examining person psychological state. With the development of computer technology, many new features were reckoned with, like aspects, or subjectivity of the text, subject and object of speech. For example, CSS solves the problem of aspect analysis by filtering dozens of messages by keywords to grant a more accurate vision of a concept.

The main method for solving the task of sentiment analysis is classification, a process of categorizing the data into classes. It could be binary (in case of sentiment analysis: positive or negative) or muti-class (in case of sentiment analysis: “0” for strongly negative, “5” for strongly positive). Usually, this task is divided into two sub-topics [23]. First is the document-level sentiment classification, the classification of the whole text by polarity. Second is the sentence-level sentiment classification, the classification of the clause as subjective or objective, and then classifying extracted subjective clause by polarity.

The main idea of this work is to update existing methods of sentiment analysis by implementing in them the sentiment framework based on the bond between predicate, subject, and object of the speech. In the work [24], Rashkin pushed the issue of connotational frames, where he examined the connection between the subject and the object and how the predicate affects their polarity. Such an approach could evaluate the opinion of an author for objects of speech. Also, the same predicate could have the opposite effect on objects of speech, e.g.“X wins Y” is good for X, and bad for Y. The positive or negative semantics for objects of speech could vary depending on the verb that was used by the author. Sentiment frames are designed to solve this problem. As the base for this research, we decide to choose the sentiment framework that was collected by Loukachevitch in [13].

To extract the relations between the objects of speech we should extract the arguments of a predicate. One of the solutions to this problem is human tagging. It is a very resource-consuming method but highly accurate. Another solution is UDPipe [25], a highly developed syntactic parser. It can mark arguments of a predicate automatically and make dependencies trees for the whole sentence.

Sentiment lexicons are an important part of sentiment analysis [1]. They represent the polarity that each word brings. Most of them are simple distribution between positive or negative polarity, but some of them could use another idea of dimensional distribution. In the EmoLex [18], words divided into three topics: Valence, Arousal, Dominance. Which differs from binary distribution of polarity or standard emotional fragmentation (anger, fear, sadness, joy). Such an approach could shed some light on this problem and could be used for the broadening of the existing lexicon.

There are several methods for creating sentiment lexicons. The first of them is human labeling: by crowdsourcing, researchers collect the data about the polarity of a word, divide the data and distribute it to a large number of annotators (EmoLex, NRC VAD Lexicon). The second method is Label Propagation [17]. It is a semi-supervised automatic method for broadening the existing small lexicon by the mode of Nearest Neighbours.

Disambiguation affects words in that way when they could reverse polarity in other topics. It entails many mistakes and noise if we don’t take it into the account (e.g. “cold” could be positive if we are talking about rooms in the south countries and negative if we are talking about food). To solve this problem most of the researchers train the model on certain topical datasets [16][17], where the semantic field of a word is statical. But such an approach disables the incorporation of these models in other tasks. We chose the Twitter corpus to evade the problem of topical lexicons and datasets, but disambiguation could still be a problem. WordNet could solve this potential problem by using the large dataset of meaning and contexts. It could easily separate “bat” like an animal and “bat” like a sports inventory.

To conclude, our aim in this work to broaden the existing sentiment lexicon and sentiment framework of Loukachevitch, using the data from the Russian segment of Twitter. Collected data will be used for improving standard methods of sentiment classification to make predictions more accurate.

## Methods

The first aim of this project is to broaden the existing sentiment lexicon and the sentiment frames from the work [13]. For this task, we will use a semantic parser UDPipe to extract new sentiment connections by building dependency trees of sentences. The extracted information will be transformed into the framework that will be embedded in the dataset before transforming text into the vector format.

After that, the next step of our work will be the representation of tweets in vector format. A number of techniques have been developed for executing this task. In this work, we will try different approaches to this task. The first of them is TF-IDF[11] (Term Frequency Inverse Document Frequency). This method based on a numerical statistic that is intended to reflect how important a word is to a document in a corpus.

The second method that we will be trying to use is Doc2Vec[10]. It is based on PV-DM (Paragraph Vector Disturbed Memory) and PV-DBOW (Paragraph Vector Disturbed Bag Of Words). PV-DM concept is to try to remember the topic or the context of the text, to make vector more representative to the document. When PV-DM is trying to predict the topic and the context by the words, PV-DBOW is trying to predict words by topic. As the output of these two algorithms, this model generates the document vector.

The third method that we will be trying to implement is LSTM networks[9]. LSTM (Long Short Term Memory) is an artificial recurrent neural network architecture. The main advantage of this network is the ability to learn long-term addictions, which makes them one of the best methods to transform texts into numerical representations. Text passes through four layers and converts into a vector that displays the main topic of the whole document.

Also, we could use in our task transformers with high performance, such as GPT-2 and BERT, but with using this method we will be faced with difficulties associated with preprocessing of our data. BERT and GPT are already trained models and it will be a difficult task to train it from the scratch. TF-IDF doesn’t seem like a decent method too. Despite its simplicity and speed, it is quite outdated and outperformed by the above methods.

## Conclusion

One of our goals is to broaden the sentiment non-topical lexicon using sentiment frames for extracting the implicit polarity of words. The main idea is to make a framework that could use data from sentiment frames and make advanced word polarity determination. This lexicon could be used as a base for future researches about sentiment analysis that uses initial sentiment data for encoding sentences into embeddings. Also in this work, we hope to realize a decent model of sentiment analysis based on the extended lexicon that we get from sentiment frames. The main problem is implementing them into the analysis model and make them accurate despite the non-topical lexicon. We think this work could be beneficial for future research. According to Sergey Smetanin and Michail Komarov SOTA on RuTweetCorp is 87.38% using M-BERT-Base-FiT[5]. We don’t set a goal to beat this result but we think that our method should outperform simple lexicon-based sentiment analysis by improving it by using sentiment frames and implementing them into the model. Even without the goal of beating SOTA, we hope that this method will achieve good accuracy to be useful for future researches.

Future works could be related to the results of this research and may be implemented in more advanced models for getting better accuracy of predictions. Also, the extended lexicon could be broadened by another method, for example adding specific words to adjusting information on the topic. This lexicon could be updated in an advanced data frame of polarized words and implemented in the word analyzing model.

## List of references (in APA format)

1. Jurafsky, D. (2000). *Speech & language processing*. Pearson Education India.
2. Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics* (p. 271). Association for Computational Linguistics.
3. Rubcova Y. Avtomaticheskoe postroenie i analiz korpusa korotkih tekstov (postov mikroblogov) dlya zadachi razrabotki i trenirovki tonovogo klassifikatora //Inzheneriya znanij i tekhnologii semanticheskogo veba. – 2012. – V. 1. – pp. 109-116.
4. Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.
5. Smetanin, S., & Komarov, M. (2019, July). Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks. In *2019 IEEE 21st Conference on Business Informatics (CBI)* (Vol. 1, pp. 482-486). IEEE.
6. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.
9. Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735-1780.
10. Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).
11. Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.
12. Deng, L., Wiebe, J., & Choi, Y. (2014, August). Joint inference and disambiguation of implicit sentiments via implicature constraints. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 79-88).
13. Карнаухова В.А., Лукашевич Н.В., Принципы создания системы оценочных фреймов для русского языка - 2018
14. Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.
15. Snyder, B., & Barzilay, R. (2007, April). Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (pp. 300-307).
16. Thongtan, T., & Phienthrakul, T. (2019, July). Sentiment Classification using Document Embeddings trained with Cosine Similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (pp. 407-414).
17. Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training.
18. Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174-184).
19. Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.
20. Hatzivassiloglou, V., & McKeown, K. R. (1997, July). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics* (pp. 174-181). Association for Computational Linguistics.
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
22. <https://blog.paralleldots.com/product/contextual-semantic-search-concept-analysis-applications/>
23. Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, *2*(2010), 627-666.
24. Rashkin, H., Singh, S., & Choi, Y. (2015). Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.
25. Straka, M., Hajic, J., & Straková, J. (2016, May). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4290-4297).