

Russian Language Sentiment Analysis

1 Introduction

2 Background

Sentiment analysis is a discipline within natural language processing, analysing people’s opinions, attitudes and emotions towards entities and their attributes, as is expressed in written text (Liu, 2015).

Here, the term entity can represent products or services, organisations, events or issues, as well as a host of other THINGS that exist within the world of which people can hold opinions and have attitudes towards.

Sentiment analysis has wide ranging commercial application, as businesses and organisations can leverage consumer opinions to make more accurate decisions regarding their product or service (Ain et al., 2017).

For organisations such as governments or local councils, sentiment analysis offers an opportunity to understand public opinion of policy, proposed or existing.

A limitation for natural language processing tasks in general is that for the abundance of English language models, curated data sets and applied research, other languages are incredibly under-represented (Bender et al., 2021; Hovy and Prabhunoye, 2021).

To analyse non-English language, there are generally two approaches.

The first, as done by Bautin, Vijayarenu, and Skiena (2008), is to use machine translation and perform the sentiment analysis on the English translation of the foreign language text.

The efficacy of such a method, clearly, depends also on the accuracy of the machine translation model. For major world languages this technique is widely applicable, but for smaller languages, it is more difficult.

The second method requires the construction of corpora in the target language. A corpus is a collection of complete and self-contained texts, which contains information consisting of emotional expressions (Peng, Cambria, and Hussain, 2017).

Analysis of population sentiment can be very useful in wartime. A study by Bourret, Wines, and Mendes (2016) analysed from social media data of inhabitants of Yemen over a three month period in 2013. The author concluded that sentiment analysis of the social media data alone could be used for evaluation negative sentiment towards violent extremist organisations or for the standing government.

Sentiment analysis of news articles has been used by social science researchers to investigate research questions related to public opinion. AP-PRAISALTHEORY

News articles are a good source for building a corpus as there is wide ranging sentiment contained within the texts, as well as generally containing a good lexical sample.

In light of recent events regarding the invasion of Ukraine by the Russian Federation, now more than ever there is a visible need for language models to determine sentiment in languages other than English.

An example where this could be applied is to measure the well-being of citizens living in areas of foreign occupation.

Further, as shown by Alonso et al. (2021), that sentiment analysis can be used as a method of detecting fake news.

3 Objectives

The following report details an attempt at sentiment analysis in the Russian language using the second method previously described. That is, by using a Russian language corpus on which to train the model.

The aim of the project is to build a sentiment analysis model that can then predict sentiment from Russian language news sources from the Russian Federation, as well as Ukraine.

The null hypothesis to be tested would therefore be that there is no significant difference in the sentiment of news articles from the two countries.

4 Methodology

The training data set to be used consists of 8,263 Russian language news articles with predetermined sentiment labels of either "positive", "negative" and "neutral". A test data set consists of 2056 samples without labels. The data was originally hosted for a Kaggle competition in 2018 (*Sentiment Analysis in Russian* n.d.).

4.1 Data Exploration

An initial exploration of the data was required to later improve model training.

Firstly, summary statistics for number of words per sample were generated. There is an average of 436 words in sample, with a maximum value of 49,039 words. The long-tail of this distribution is visualised in figure 1.

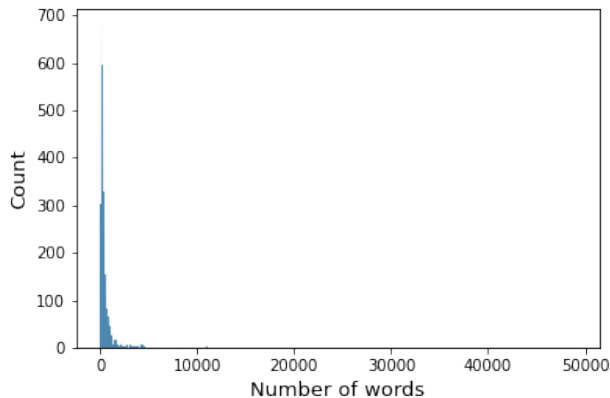


Figure 1: Distribution of the length of texts.

This means that without any kind of augmentation, the overwhelming majority of samples would need to be padded by two orders of magnitude, vastly increasing the size of input data for the model. A maximum sample length of 500 words was chosen as a filter, leading to training and test data sets of 6284 and 1563 samples, or 76.05% and 76.02% of the original data set respectively.

As can be seen in figure 2, after generating this subset of the data, the new mean number of words per sample is 209.

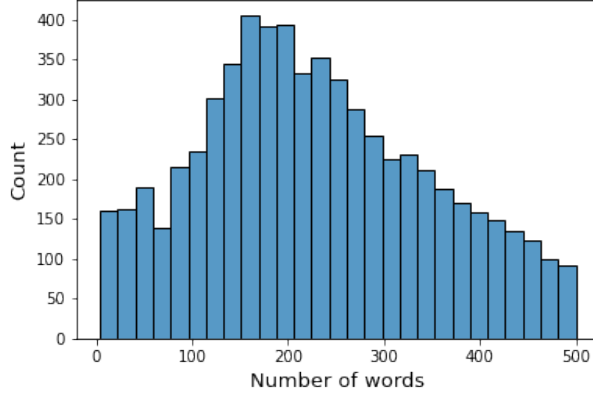


Figure 2: Distribution of the length of texts (500 word maximum).

An analysis of the entire data set at the word level showed a distinct corpus of 290,028 words, and for the subset 135,779, reducing the total unique word count to 46.82% of the original data.

4.2 Preprocessing of Data

In order to prepare the data as an input to the neural networks, it was first converted to lower-case and stripped of punctuation, numbers and special characters.

Next, a collection of Russian stopwords were removed from the samples. (Romanov, Lomotin, and Kozolva, 2019)

A final step was taken to stem the words in the samples. This was achieved by use of a Snowball Stemmer by the Natural Language Toolkit package (Bird, Klein, and Loper, 2009).

Other preprocessing steps that were initially included but were found to not increase the performance of models were the removal of the most frequent words and the least common words.

The summary statistics for the data were then repeated for the reduced cleaned data set, leading to the following observations.

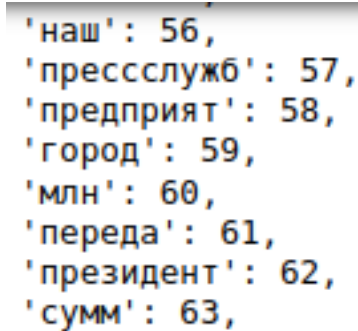
After filtering again for samples within the limit of 500 words, the number of training examples was 6624, and test 1664, or 80.16% and 80.93% of full cleaned data set respectively.

A total of 69.503 unique distinct words existed in the full cleaned data

set, with 37,420 in the cleaned data set with the 500 word constraint.

Hence the ratio of words was 53.84%, which is higher than that found in the original data set. This is to be expected as many rarely used conjugations or cases for particular words will have been retained given their stem is more common.

Once the textual data was in a cleaned state, it was necessary to integer-encode the samples at the word-level. That is, a word-level integer-encoder was used that associates a unique integer value to every unique word in the text (Figure 3).



```
'наш': 56,  
'прессслужб': 57,  
'предприят': 58,  
'город': 59,  
'млн': 60,  
'переда': 61,  
'президент': 62,  
'сумм': 63,
```

Figure 3: Integer encoding of Snowball-stemmed vocabulary.

The final step was to pad the samples so that they were all the same length as neural networks require fixed-dimensional tensor inputs. Each sample was padded with zeros after the end of the text so that it could be represented as a one-dimensional 500-element array.

4.3 Model Architecture

5 Experiments

For all neural network experiments undertaken in this project, a number of parameters have been kept constant.

Validation data has been extracted from the training set in a shuffled and stratified manner, with a size of 20% of the total training set.

All models have been trained using the full vocabulary found in the training set, with a 30-dimensional embedding vector.

5.1 Hyperparameter Selection

5.2 Word Embeddings

5.3 Transfer-Learning

6 Results

7 Conclusion

References

- Bing Liu (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.
- Qurat Tul Ain et al. (2017). “Sentiment analysis using deep learning techniques: a review”. In: *Int J Adv Comput Sci Appl* 8.6, p. 424.
- Emily M. Bender et al. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. Virtual Event, Canada: Association for Computing Machinery, 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: <https://doi.org/10.1145/3442188.3445922>.
- Dirk Hovy and Shrimai Prabhumoye (2021). “Five sources of bias in natural language processing”. In: *Language and Linguistics Compass* 15.8, e12432.
- Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena (2008). “International sentiment analysis for news and blogs”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 2. 1, pp. 19–26.
- Haiyun Peng, Erik Cambria, and Amir Hussain (2017). “A review of sentiment analysis research in Chinese language”. In: *Cognitive Computation* 9.4, pp. 423–435.
- Andrew K. Bourret, Joshua D. Wines, and Jason M. Mendes (2016). “Assessing Sentiment in Conflict Zones Through Social Media”. PhD thesis. Naval Postgraduate School, Monterey, California.
- Miguel A Alonso et al. (2021). “Sentiment analysis for fake news detection”. In: *Electronics* 10.11, p. 1348.

- Sentiment Analysis in Russian* (n.d.). <https://www.kaggle.com/competitions/sentiment-analysis-in-russian/data>. Accessed: 2022-02-20.
- Aleksandr Romanov, Konstantin Lomotin, and Ekaterina Kozolva (2019). “Application of Natural Language Processing Algorithms to the Task of Automatic Classification of Russian Scientific Texts”. In: *Data Science Journal* 18.1, pp. 1–17.
- Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.