Технические науки

UDC 004.896

**Zhunissov Darkhan**
*Master Student of the*
*International Information Technology University*

**Aliaskarov Serik**
*PhD Student of the*
*International Information Technology University*

**Satybaldiyeva Ryskhan**
*PhD, Associate Professor*
*International Information Technology University*

# METHOD OF TEXT SENTIMENT ANALYSIS USING NEURAL NETWORKS

***Summary.*** *In this work, methods of calculating and analyzing the method for determining the sentiment of texts were described. And also, the main parameters for the application of neural networks for natural language processing were identified. For this work, some implementations of neural network algorithms were analyzed and shown. Algorithms for processing text data were used to develop a system for analyzing the sentiment of texts, as well as the results of a study of a natural language processing system.*

***Key words:*** *processing, natural languages, classification, analysis, categorization, recognition, understanding.*

***Аннотация.*** *В данной работе были описаны методы расчёта и анализа способа определения тональности слов. А еще, были обнаружены главные характеристики для использования нейронных сетей для обработки естественного языков. Для данной работы были* *проанализированы и показаны кое-какие реализации алгоритмов нейронных сетей. Для разработки системы системы по анализу тональности слов были использованы методы по обработке текстовых данных, а еще итоги изучения системы обработки естественного языка.*

***Ключевые слова:*** *обработка, естественные языки, классификация, анализ, категоризация, распознавание, понимание.*

**Introduction.** In the digital world, the importance of information processing becomes more important every day. In connection with the rapid development of technologies and the development of systems for predicting social grids, one of the main directions is considered to be systems for processing natural language. Natural language processing systems have every chance to cultivate all sorts of forms of textual information as news publications, for example, notes and comments for them. An unpretentious look through the eyes of buyers of systems, this is presented in the guise of targeted services, the formation of a portrait of the user and the forecasting of actions based on the awareness of words and actions. As you know, one of the most necessary technologies is considered to be methods based on NLP (natural language processing), for example, as of today, the provided area is not sufficiently researched. Natural language processing is an area that emerged as a result of the synchronization of sciences such as linguistics and mathematics [1]. However, systematization alone is not enough for a clear understanding of the word; you need to categorize the word for further processing. Classification of words is a process of analytical and syntactic analysis of words in various fields for subsequent work with data [2]. Next comes in the footsteps of data visualization and preprocessing for making conclusions in information systems. However, for the systematization of words, clear methods of machine learning and gigantic corpuses of labeled data are essential, which make the risks of the costs of the semantic meaning of categories. As a consequence, for word processing in particular for determining the sentiment of words, a more efficient way of processing and analyzing words will be required. The introduction of more comprehensive and difficult word analysis methods has the potential to reduce the possibility of misses, but will entail huge resource demands. It is expected, in fact, that there is a need to increase the accuracy and the introduction of more advanced algorithms for systematization.

**Methodology.** This article will consider the methods of text classification, therefore the object of the research is text analysis. There are no standardized templates for "text analysis". Text analysis is an example of an area of interdisciplinary research, respectively, the concept of analysis is transformed depending on the area and on what is the general context of the article associated with this concept.

Sentiment analysis is a fairly common task in machine learning and is explored in many areas, as well as problems posed by science. Machine learning specialists working on complex algorithms of neural networks strive to ensure their necessary functioning in certain, previously unknown conditions, that is, they achieve the necessary forms of behavior from the systems and algorithms being developed. Linguists study the information that determines the composite functioning of the text, as well as the unit of each word in the right context. The key functions of cognitive linguistics are the features of assimilation and processing of information [3]. Relatively speaking, the processing of textual information is the finding of sequences in the text, received from various informational articles. For this article, the above description is the definition of information processing, that is, considering the text as a unit of calculation, and the classification of the text is, then - how it shows itself in different conditions. Although one definition is not enough, it is also necessary to identify measurable and measurable measures, and in this case it is the context in which the units participate for further preprocessing.

Within the framework of this study, the method of determining the sentiment acts as the basis for further interactions in the system, complementing the system as a full-fledged cycle of text comprehension. Therefore, the primary action that determines the category of the text is segmentation. Each unit of text will have a certain weight and will be considered within the entire context, which will build links.

The corpus of data for the provided study was obtained by uploading the corpus of big data from notes on Wikipedia and data in the public domain. A total of 600,000 original texts were received. For the successful implementation of the provided task, the data corpus was expanded by the method of researching public networks to unload the largest number of original phrases and sentences. The continuous content of the notes was carried out by the method of research and identification of the main topics, by the method of adding the primary corpus of research tags. Research tags were set for searches on public networks and news portals. The role of the man in the collection was minimal. By evaluating the uniqueness of the proposals, they were added to the big data corpus and created clusters of original services. The bulk of the words were in Russian and Kazakh, in fact, which made it superior for collecting data in Kazakh and Russian by tags. All data sets were not formatted or processed during the collection process, and the corpus had the opportunity to get stop words and insignificant texts. In the joint difficulty, programs were used to collect from 12 social networks, and another 800 news portals, in fact, which gave a great exaggeration in the quality and number of the corpus for further processing.

To collect data from open sources, the search-key technique was used, which works on the basis of the python language and tools for unloading the html version of the page and finding by tags. Tags were applied in accordance with the theme and were obtained by searching within sources, viewing tags and searching through search engines.

To collect data from news portals, a search by topic and further by tags was applied. The portal was unloaded by a separate collector script through specific periods of time and saved to the corpus for further processing. Word searches were carried out on news announcements and then on descriptions under the news, this was implemented through the methods of reconnaissance on the website and the search engine by filtering the time marker from new to old.

A total of 21,231 texts were collected from social media and news portals. For example, the search, by tags, was on political and social topics, statistics highlight a portrait of the mood of the population in the time frame of the search.

Sentiment analysis by type was broken down into a number of milestones, they consist of functions and basic tags. When dividing into tones, the concept of bordering components occurred. Subsequently, the percentage correspondence of proximity to a particular class was revealed. Linguistic characteristics were different in all and were analyzed.

For automatic classification, the division was by means of instruments, and the instruments gave out a category, closer to the word, for addition.

Models and layers were applied from keras library. Keras is an open source neural network library written in Python [4]. The goal of Keras is considered to be operational deep learning. For the model, the text was removed from the original and returned to its original form in the form of an array of phrases, and then texts. Keras pad_sequences was used to ensure that all the queues in the list are of the same length. By default, this is done by adding 0 at the beginning of any sequence, until any sequence will have the same length as the sequence alone. To identify the sentiment, the implementation of the LSTM recurrent neural network was applied. Weight training allows you to learn the function that minimizes costs.

As part of the implementation of our tool for organizing into categories, we made our classifier. For this, techniques were applied to remove stop words from a word, substitute letters in lowercase letters and drive the original form of texts. Next, the Tokenizer function from the keras library was used. Tokenizer converted words in order and translated them into vector representation as numeric. Further, with support for pad_sequences from the keras preprocessing library, we returned queues with intervals and labels.

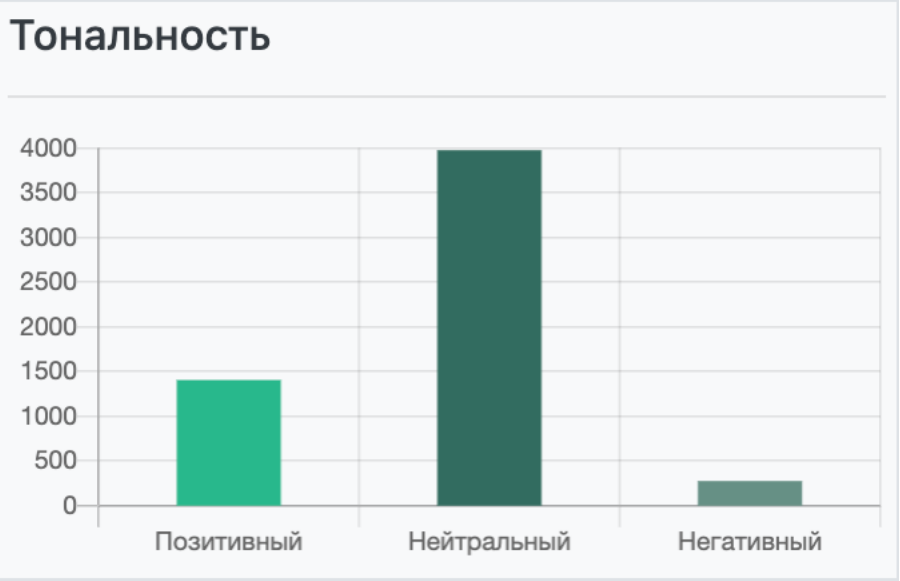
To conclude the difficulty of systematizing the determination of the sentiment, we ran the acquired results from our classifier for working with input data into the predict function of our model built on the basis of LSTM (Long short Term Memory). LSTM is one of the variations of classifiers based on the RNN which is used mainly in problems of processing polynomially distributed data as a text classification or image classification [5]. To optimize the workflow, we pre-saved selections of the model training result in h5 format and loaded the result through the keras library function load_model. In the solution, the result was obtained, with the support of indexing the total, the category was displayed. The first results of categorization can be seen in Table 1, and shows the accuracy of the model about 82%.

*Table 1*

**The first result of the categorization of texts on the example of a large data corpus**

|  | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.83 | 0.8 | 2777 |
| 1 | 0.82 | 0.77 | 0.8 | 2763 |
|  |  |  |  |  |
| micro-avg | 0.8 | 0.8 | 0.8 | 5540 |
| macro-avg | 0.8 | 0.8 | 0.8 | 5540 |
| weighted-avg | 0.8 | 0.8 | 0.8 | 5540 |

In order to handle different cases of input, we can collect data from different sources and areas of activity, so the quality of the collected data set will increase, which will have a positive effect on the operation of the model. In the process of implementing the model, we increased the accuracy of the model using logistic regression and the accuracy result increased to 84%. The method will work well on various types of data, including irony and satire, as there is constant improvement due to the quantity and quality of the data.



**Pic. 2. The result of determining the sentiment of texts on the example of a large data corpus after model improvements**

Like all other systems, theoretically, a classification system based on text analysis can be deceived if the texts do not come in prepared templates, but there is a percentage of error for this. Despite this, in practice, the system works stably and shows good results in the classification of texts. Texts coming from information systems, as well as from various sources, are classified by sentiment and provide an opportunity for subsequent processing for understanding and analyzing natural language. This work is based on the hypothesis that the texts are unstructured and collected from different sources, but there is a percentage of error in identifying tones depending on the area of the context itself. The solution to this problem is to increase the data corpus, train models for the necessary ones, and work on cleaning the data and identifying key parameters.

**Conclusion.** In conclusion, the purpose of determining the tone of a text based on its features is to ensure a holistic understanding and processing of natural languages. You can analyze text only by comparing data received from units of text that are part of the context. Since the general context plays a key role in determining the sentiment of the text, it is necessary to take into account the peculiarities of the linguistic corpus, tokenization of words and reduction to their original form. In order for there to be no deviations in the accuracy of the results, it is necessary to work on the cleanliness of the case and to give significant attention to the elaboration of the model and the alignment of the parameters. Thus, this model will have a percentage of error in the calculations, but it will constantly self-learn and give better results every time, since the text corpus will need to be enlarged, and attention will be paid to cleaning.

## References

1. Bol'shakova E.I., Voroncov K.V., Efremova N.E., Klyshinskij E.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh [Automatic Natural Language Processing and Data Analysis] NIU VSHE, Moscow, 2017. P. 124-128.
2. Rafanov S.M. K probleme klassifikacii tekstov v mashinnom perevode [the problem of text classification in machine translation], KRSU, Moscow, 2013. P. 36-42.
3. Popova Z.D., Kognitivnaya Lingvistika [Cognitive Linguistics], Federal'noe agentstvo po obrazovaniyu Voronezhskij Gosudarstvennyj Universitet, Voronezh, 2007. P. 7-12.
4. Ketkar Nikhil Introduction to Keras. 2017. DOI: 10.1007/978-1-4842-2766-4_7.
5. Xu Shuo & Li Yan & Zheng Wang Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 2017. P. 347-352. DOI: 10.1007/978-981-10-5041-1_57.