Sentiment analysis of multilingual texts using machine learning methods

Anton I. Kanev
Chair of Information Processing and Management Systems
Bauman Moscow State Technical University
Moscow, Russia
aikanev@bmstu.ru

Grigory A. Savchenko
Chair of Information Processing and Management Systems
Bauman Moscow State Technical University
Moscow, Russia
sgfox4@gmail.com

Ilya A. Grishin
Chair of Information Processing and Management Systems
Bauman Moscow State Technical University
Moscow, Russia
ilia-grishin@mail.ru

Denis A. Vasiliev
Chair of Information Processing and Management Systems
Bauman Moscow State Technical University
Moscow, Russia
tffoem@gmail.com

Emilia M. Duma
Chair of Information Processing and Management Systems
Bauman Moscow State Technical University
Moscow, Russia
em.duma.2001@mail.ru

***Abstract*— Sentiment analysis is a topical task of evaluating the content of texts, articles and statements in order to study public opinion and the relationship between the moods of users and their phrases. At the same time, the analysis of multilingual texts is more difficult. Therefore, in this work a study was carried out with datasets in different languages, including the use of automatic translation. Various machine learning methods, several neural network architectures, and the VADER analyzer were applied. Also, NER was combined with other techniques in the work to determine the sentiment of individual entities. The authors evaluated the methods on various datasets using the F-measure. The obtained metric values show the best results for a neural network.**

***Keywords*— Nature language processing, sentiment analysis, machine learning, multilingual text**

I. INTRODUCTION

A. Sentiment Analysis

Sentiment analysis is a class of content analysis methods included in natural language processing (NLP) tools and representing the understanding of language in the area of computer analysis, the main task of which is to classify emotions, determine the emotional coloring of the text and identify evaluative vocabulary in relation to the objects of assessment.

Most often, the tonality of feelings and emotions is divided into positive, neutral and negative. However, the variety of turns of speech when evaluating objects or expressions of opinion in many languages, and in particular in Russian, is so extensive that sometimes even a person cannot determine exactly which of the three categories of sentiment indicated above can be attributed to a statement, but only probabilistically evaluate the attitude the statement in question to categories. For example, journalists and news agencies, trying to observe objectivity in their statements and avoiding a direct assessment of the events that have taken place, do not use positive and negative sentiments within the area of their profession, but most often resort to neutral [1].

However, this problem can be useful in some tasks, therefore, analysts and researchers, for the tasks of recognizing changes in mood intensity over time, use the power of the mood of the statement based on valence, and not the so-called binary polarity (positive or negative tonality) [2].

If the sentiment analysis for the English language is already well developed, the number of works devoted to the analysis of the sentiment of Russian-language texts has been increasing only since 2014. Up to 2017, the number of such studies is growing rapidly. The main directions of the analysis of the emotional coloring of texts are approaches based on machine learning, approaches based on lexical and semantic rules and classification rules, approaches based on neural network architectures and cloud services, which are hybrid solutions.

With the emergence of a wide number of social networks, forums and Internet sites, an abundance of data sets of various subject areas and their research appears, in order to study public opinion, patterns, dependencies and the relationship between user moods and their words or phrases.

Thus, sentiment analysis is used for a wide range of subject areas: reviews on laptops, hotels and books [3], determining satisfaction with the quality of animal images and services of the national park [4], determining the quality of applications based on reviews on mobile applications [5], analysis of user sentiments for large-scale events and movements in their country [6], analysis of toxic comments for a chatbot [7], determining the mood and emotions of the general public about COVID-19 [8].

B. Current Investigations

Currently, most research in sentiment analysis can be divided into several categories depending on the area of study: user generated content from social media, product and service reviews, news from the media, books, and mixed data sources [9].

Among the data from social networks, comments left by users are most often analyzed, for example, on Twitter [10, 11, 12], YouTube [13]. Some authors investigate more specific cases, for example, in the article [14] they consider github issues.

978-1-6654-0993-3/22/$31.00 ©2022 IEEE

There are several approaches to text sentiment analysis that are applicable to data in Russian. This is a rule-based approach, where frequently occurring words or phrases are selected from the text, which can be assigned a positive or negative assessment, a machine learning approach, where a classifier is built based on a training, pre-labeled collection of texts and hybrid approaches. including both the advantages and disadvantages of the previous ones.

In [15], good results were shown when using CNN (Convolutional Neural Network), and the inclusion of emoji in the analysis improved the accuracy of the F-measure from 74.31% to 75.45%. In other cases, as, for example, in articles [16, 17], the use of RNN (Recurrent neural network) also gives good results. RNN trained on a combination of bag-of-words (BOW) and wor2vec presented an advantage over state-of-art approaches, F-measure improved from 78.12% to 85.01%.

Similarly, with the previous ones, other machine learning methods are used (Naïve Bayes Classification method, Support Vector Machine Classification Method, Maximum Entropy Classification method) [12].

Special tools have been developed that contain a whole range of methods for solving NLP problems. One of them is VADER (Valence Aware Dictionary and sEntiment Reasoner), developed for sentiment analysis of data obtained from social networks, which was used, for example, to determine the best voice assistant in the opinion of English-speaking users. [17].

Another interesting tool is Natasha. Initially, this tool only solved the Named Entity Recognition (NER) problem for the Russian language. At the moment it is a large project that combines 9 different tools for solving NLP problems.

Previously, NER has already been used in combination with the sentiment analysis tool to create a chat bot that analyzes and processes information in the Indian language [7]. We tried using Natasha to solve the sentiment analysis problem.

When analyzing texts in Russian, the problem of the language itself is that it is rich in polysemous and similar in meaning. The authors of the article [18] have shown that the addition of the initial data with synonyms, adjectives or verbs that match the existing words can have a positive effect on the result. Thus, the addition of adjectives improved the accuracy measure from 84.57% to 87%.

II. Methods

At the moment, there is a large number of works on machine learning and deep learning, but these works predominantly use monolingual texts. Therefore, in this work, the authors propose to compare machine learning methods in different languages, and apply NER to improve the quality of determining the emotional coloring of texts.

A. TF-IDF

Initially, a text is prepared for sentiment analysis. TF-IDF (Term Frequency times Inverse Document Frequency) is a measure of the importance of a word, shows how often a word occurs in a set of documents (1). One word in each of the N documents occurs a different number of times. TF_{ij} - Term Frequency times - a measure of the frequency of occurrence of a word in a document. IDF_i - Inverse Document Frequency - a measure of the inverse frequency of a document. The words that appear most often in documents and have the largest TF-IDF, as a rule, describe the main topic of discussion in a particular document.

$$TF\text{-}IDF=TF_{ij} \cdot IDF_i \quad (1)$$

Fixed-length numeric vectors are called embeddings. They describe a specific entity and are used for NLP, natural language processing. A numerical vector of length k is a series of k numbers, in which they stand in a strict, definite order. Due to embeddings, machine learning methods determine the relationship of certain words to each other. For example, the words "man-woman" are connected by the difference vector between them. If we add the word “king” to the word “woman”, then the word “queen” will turn out. a woman with a crown on her head is a queen.

B. Machine Learning

In our work, neural networks solve classification and regression problems. In short: the classification task is to predict the label, and the regression task is to predict the quantity. Consider two types of regression: linear and logistic. Linear regression shows a linear relationship between two or more variables. Linear regression is used to find the line of best fit. This is the line along which the largest number of values are located. Used for continuous values such as height or weight.

Logistic regression is different from linear regression. By analyzing the dependent variable and several independent ones, it can predict the occurrence of a certain event. The answer is given in binary form - 0 or 1. The dependent variable in linear regression is continuous, and in logistic regression the variable has a limited number of values. The linear regression (2) with parameters m and c looks like:

$$y=mx+c \quad (2)$$

Meanwhile, logistic regression (3) differs:

$$y=e^x+e^{-x} \quad (3)$$

The methods for dealing with errors in these regressions are different: the linear one cuts out the errors with squares, the method of minimizing the error by the least squares of the model to the data is used, the logistic one removes the error of the asymptotic constant, the function of logistic losses is used. From this we can conclude that logistic regression solves its problem better because it does not cut out the correct points, unlike linear regression with its Least Squares method.

Support vector machine is a popular machine learning method that solves regression and classification problems. This method builds planes between objects of different classes and looks at the size of the gap. The gap is inversely related to the average error: the gap is smaller - the error is larger, the gap is larger - the error is smaller.

This method outperforms stochastic descent and neural networks in several respects: it is similar to a two-layer neural network, the number of neurons and support vectors is determined automatically. It makes the classification more confidently because the dividing hyperplane maximizes the width of the dividing strip. The disadvantages of this method are: features are not selected, there are no descriptions of

methods for constructing kernels and straightening spaces, the constant is selected using cross-validation.

Ensemble methods are used to improve the accuracy of machine learning results. A feature of these methods is the training of several models to solve one problem, and in the future, the combination of the results. The end result is the best possible result. Decision trees are popular reference models for ensemble methods. Weak trees are shallow trees (few nodes) or deep trees (many nodes). Shallow trees have less spread, but they have higher offset. The opposite is true for deep trees.

In our work, we used the Random Forest, one of the machine learning algorithms related to bagging, consists of deep trees. At the moment, heuristics have not been able to significantly improve this algorithm. Its peculiarity lies in the fact that it consists of a large number of decision trees. These trees are combined to produce the smallest scatter result. In each node, the data received at the input is divided according to some criterion. If the data matches the given criteria, they are lowered along the Yes branch, otherwise, along the No.

C. Deep Learning

Recurrent neural networks (RNN) are often used in NLP, natural language processing, they are good at generating data flow such as sentences. They are ideal for text and speech analysis: a neural network builds inter-element connections that form a directional sequence. A feature of recurrent neural networks is the sequential use of information, that is, to predict the next word in a sentence, the network considers the previous ones (Fig. 1). This means that the order in which the data is fed to train the network becomes important.

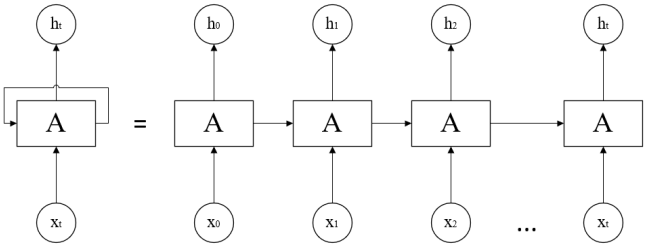

Fig. 1. Expanded recurrent neural network with RNN.

The problem of mathematical optimization is solving by an algorithm based on the gradient of a function. Stochastic Gradient Descent is an optimization algorithm mainly used to tune machine learning parameters. Batch gradient descent counts the gradient of the optimized function as the sum of the gradients from each element. Stochastic counts from a random element. Stochastic gradient descent has advantages over batch: it does not iterate over all the elements in the training set, unlike batch. Hence stochastic gradient descent finishes faster.

To determine the quality of the algorithm, we used the F1-measure. It allows us to simultaneously evaluate accuracy and completeness. The advantage of this measure is a more uniform characteristic of the model under consideration than sensitivity, specificity and precision. The downside is that True Negative is not taken into account, the metric may show an incorrect value in the case of the same accuracy and completeness. F1-measure is the quotient of the doubled product of accuracy and completeness by their sum.

D. NLTK and NER

There are two popular libraries used for sentiment analysis: VADER and TextBlob. These libraries do the analysis in virtually the same way: both assess the sentiment of a phrase. Polarity is indicated by a floating point digit, from [-1] to [+1]. [-1] is the most negative, [+1] is the most positive. There are also some differences. Textblob, in addition to assessing polarity, provides an assessment of subjectivity. VADER in phrases understands negation, denial reinforced by additional words, punctuation in the form of several exclamation marks to enhance emotionality, various slangs and acronyms.

One of the most popular NLP challenges is Named entity recognition. This process consists of two steps: discover that the word sequence is a named entity and determine which class it belongs to. Named entities come in different lengths, from one word to several. Classes - location, person, date, etc. The main difficulty in solving the NER problem is the large number of homonyms.

As an example, consider the word "Washington": is it a person or a location? There may also appear borderline cases when certain words need to be included in the entity's span or not: "Welcome to the Real Anglers Store" - here you need to add the word store to the named entity; "Agricultural store" Gardener "invites you to a closed sale" - it is not clear here if the first two words are included in the name of the store. We used NER with sentiment analysis, which allowed us to determine in what sentiment certain named entities were used.

III. Investigation

A. Datasets and analysis

Datasets in Russian and English were used in the work. Automatically annotated corpuses of tweets (RuTweetCorp [19]) were selected as datasets in Russian, containing about 220,000 positive and negative entries from November 2013 to February 2014, and reviews on clothes (RuReviews [15]), containing 30,000 entries for each mood category. Sentiment140 [20] and Twitter Sentiment Dataset [21] were selected as corpuses in English, containing 1.6 million and 160,000 data.

The paper considered various options for recognizing the sentiment of the text. Thus, a sentiment analysis was carried out using the extraction of named entities. The datasets in Russian were processed using the Natasha library package.

It was found that the use of names is associated with the mention of other users, usually in a positive context, in addition, retweets are more often used for positive tweets than in a negative context. Thus, the named entities in positive tweets were associated with the victories of football clubs (Zenit), with the proposal to abolish the visa (Russia, EU, Ukraine), with the release of new cartoons (Ivan Tsarevich and the Gray Wolf). The negative sentiment of the posts of users is observed in connection with negative news: death (Paul Walker) and accidents of celebrities (Schumacher), terrorist attacks (Volgograd), cancellation of New Year's fireworks (St. Petersburg). It is noticed that obscene expressions prevail in recordings with a negative connotation.

It can be seen that in the dataset, the entities that have been obtained are mainly associated with either constant mention in a positive or negative way, or refer to a specific time period in which a positive or negative event occurred.

Looking at the clothing reviews dataset, the following entities are seen in a positive context: cities (Moscow, St. Petersburg, Novosibirsk, Krasnoyarsk), clothing sizes (M, XL, XXL), and actions (Ordering). For example, the entity "Ordering" is usually associated with the fact that the customer liked the item and ordered it several times.

For a negative context, the following entities are characters: countries (Russia, China, Latvia), payment methods (PayPal), platforms (Aliexpress, Avito), delivery methods (Russian Post, China post registered air mail, CDEK), reasons for refusal (small size, disappointed, waited). For example, reviews that the clothes are small in size means that the item is not suitable for the buyer due to problems with the size, and therefore the buyer is unhappy with this.

It can be noticed that not all words highlighted with Natasha's NER are named entities, which reveals some errors in the definition of entities. According to the results of the study with the help of NER, we observe the erroneous assignment of reactions to sad events to negative moods. Also, named entities such as the above countries and cities were ambiguously divided into different categories. Thus, the study of the possibility of using NER to improve the quality of text analysis did not give the expected result.

B. NLTK

Sentiment analysis of datasets was carried out using the NLTK library package. This package does not support the Russian language, therefore, in this work, translation using the EasyNMT library and the Opus-MT model is used to classify the sentiment in the VADER analyzer included in the NLTK package. For all datasets, F1 metrics were obtained using the VADER sentiment analyzer (Table I).

TABLE I. SENTIMENT ANALYSIS RESULTS FOR VADER

Dataset	F-measure
RuTweetCorp	0.501
RuReviews	0.733
Twitter Sentiment Dataset	0.730
Sentiment140	0.563

The low quality of the tweet classification may be due to the low quality of the text itself, the predominance of the spoken genre in it, the use of youth jargon typical for social networks, as well as the use of abbreviations. This is especially noticeable for tweets in Russian, since the text may be incorrectly translated due to the aforementioned points, and the result may lose some details or acquire new ones.

The best quality was shown by reviews about clothes in Russian. This is due to the fact that reviews about clothes, compared to tweets, do not contain abbreviations and jargons, and they also formulate sentences and their meaning better in them.

When looking at the Twitter Sentiment Dataset, it was noticed that the sentences in it are quite clearly and qualitatively formulated, which distinguishes it from other datasets containing tweets, the superiority of the dataset is also the original text in English.

C. Machine Learning methods

The analysis of the sentiment of the text using machine learning. The following methods were used for machine learning analysis: linear regression, logistic regression, support vector machine, stochastic gradient descent, random forest.

For training, the data was normalized into a matrix of token counters using the CountVectorizer module. To do this, it composes tokens for all words, and then for each text constructs a vector, which contains the number of times each word is used in the given text.

Russian datasets were additionally translated into English (RuTweetCorp ENG and RuReviews ENG) and also processed using machine learning. After training, the following F1-measure results were obtained (Fig. 2).

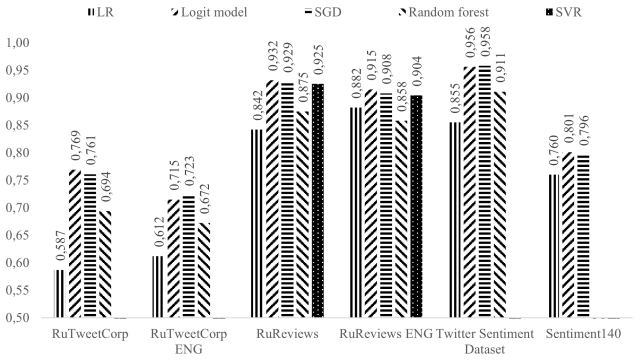

Fig. 2. Sentiment Analysis Results for Machine Learning Methods.

The absence of data on the histogram means that no results were obtained for this case. Translated data show close results to original dataset for all machine learning methods.

Support vector machine is one of the longest-trained methods, as it scales very poorly, as a result of which it does not work well with large amounts of data. The training results are roughly proportional to the results obtained with VADER. The best results are shown by logistic regression and stochastic gradient descent.

D. Recurrent Neural Network

We also analyzed the sentiment of the text using a recurrent neural network, since the use of RNN can achieve a higher recognition ability by including information about the sequence of words.

The neural network contains the following layers: embedding layer; LSTM layer; an output layer that maps the outputs of the LSTM layer to the size of the output vector; an activation layer that turns all values in the output layer into numbers from 0 to 1.

The embedding layer is necessary due to the large number of unique words used, which reaches a value of 850,000 words. Coding that many words is an extremely

inefficient task. For this, a lookup table will be used, which stores attachments of a fixed dictionary and size.

The data then enters the LSTM layer, in which cells add duplicate connections to the network and enable word sequence information to be included. The data then goes to the sigmoid output layer, where the values are converted to numbers from 0 to 1.

Training was carried out with different RNN parameters. The results of the F-measure were obtained for different batch size (Fig. 3), that shows similar values for all sizes.

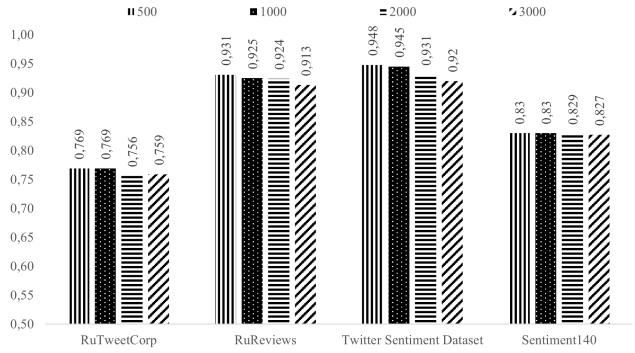

Fig. 3. Results of the F-measure for different batch size.

On the contrary, hyperparameter of learning rate (Fig. 4) affects the quality of learning process. Better results were obtained for 0.01 value.

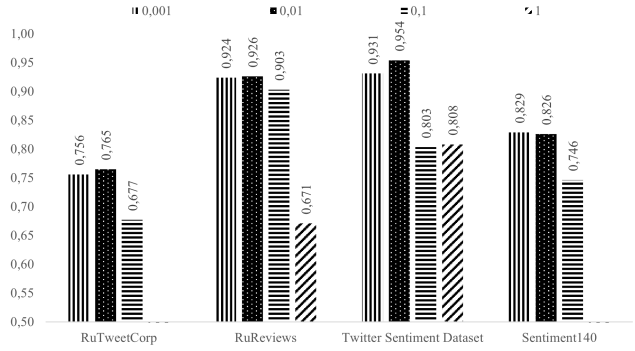

Fig. 4. Results of the F-measure for different learning rate.

Different types of optimizers (Fig. 5) were compared. They show that Adam optimizer has the higher results to SDG, Adagrad and Adadelta optimizers.

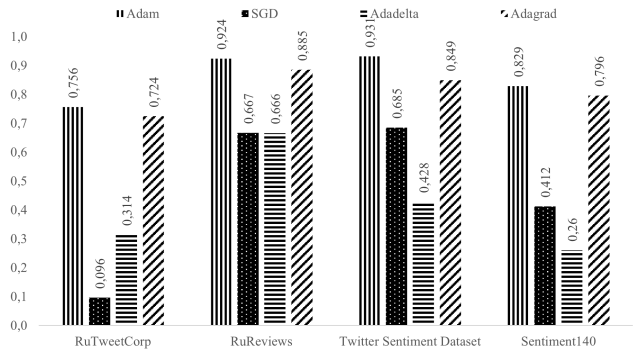

Fig. 5. Results of the F-measure for different optimizers.

GRU and LSTM cells performs sentiment analysis with close values of F-measure (Fig. 6). Therefore, it is better to use GRU cells with fewer number of parameters.

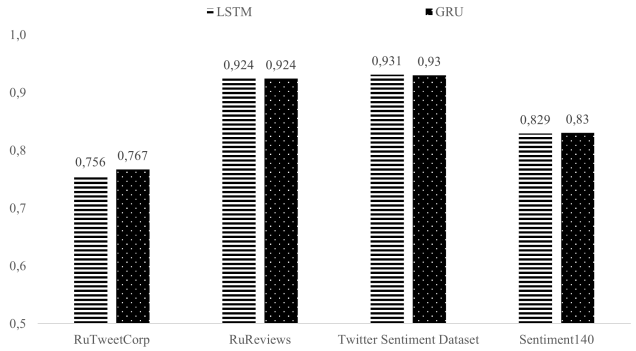

Fig. 6. Results of the F-measure for different classification layers.

It can be seen that RNN has shown results that are comparable to logistic regression method. Thus, the results obtained are comparable with the results of the study [8], where using the Bidirectional Gated Recurrent Unit on the RuTweetCorp and RuSentiment datasets, F1-metrics were obtained equal to 0.91 and 0.77, respectively.

IV. Conclusion

The work investigated different methods of machine learning and recurrent neural network for sentiment analysis, but the highest indicators are for logistic regression and neural network. The authors found a slight decrease in the accuracy of the analysis using machine translation and propose it for multilingual texts. The NER results on the considered datasets showed a small number of identified entities in the texts and the difficulties of using it for sentiment analysis. In the future, the authors plan to consider texts in other languages, as well as process emoji.

References

[1] Balahur A., Steinberger R., Kabadjov M., Zavarella V., van der Goot E., Matina Halkia M. , Pouliquen B., Belyaeva J. Sentiment Analysis in the News // Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 May 2010.

[2] Wilson T., Wiebe J., Hwa R. Just How Mad Are You? Finding Strong and Weak Opinion Clauses // AAAI, 25 July 2004.

[3] Ge H., Zheng Sh., Wang Q. Based BERT-BiLSTM-ATT Model of Commodity Commentary on The Emotional Tendency Analysis // 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI).

[4] Woldemariam Y. Sentiment analysis in a cross-media analysis framework // 2016 IEEE International Conference on Big Data Analysis (ICBDA).

[5] Fan X., Li X., Du F., Li X., Wei M. Apply word vectors for sentiment analysis of APP reviews // 2016 3rd International Conference on Systems and Informatics (ICSAI).

[6] Tan X., Zhuang M., Lu X., Mao T. An Analysis of the Emotional Evolution of Large-Scale Internet Public Opinion Events Based on the BERT-LDA Hybrid Model // IEEE Access (Volume: 9).

[7] Murali S.R., Rangreji S., Vinay S., Srinivasa G. Automated NER, Sentiment Analysis and Toxic Comment Classification for a Goal-Oriented Chatbot // 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS).

[8] Nair A.J., G V., Vinayak A. Comparative study of Twitter Sentiment On COVID - 19 Tweets // 2021 5th International Conference on Computing Methodologies and Communication (ICCMC).

[9] Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives // IEEE Access (Volume: 8).

[10] Jafarian H., Taghavi A. H., Javaheri A., Rawassizadeh R. Exploiting BERT to Improve Aspect-Based Sentiment Analysis Performance on Persian Language // 2021 7th International Conference on Web Research (ICWR).

[11] Wongkar M., Angdresey A. Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter // 2019 Fourth International Conference on Informatics and Computing (ICIC).

[12] Mandloi L., Patel R. Twitter Sentiments Analysis Using Machine Learninig Methods // 2020 International Conference for Emerging Technology (INCET).

[13] Singh Sh., Sikka G. YouTube Sentiment Analysis on US Elections 2020 // 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC).

[14] Ding J., Sun H., Wang X., Liu X. Entity-Level Sentiment Analysis of Issue Comments // 2018 IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion).

[15] Smetanin S., Komarov M. Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks // 2019 IEEE 21st Conference on Business Informatics (CBI).

[16] Cheng J., Sadiq M., Kalugina O.A., Nafees S., Umer Q. Convolutional Neural Network Based Approval Prediction of Enhancement Reports // IEEE Access (Volume: 9).

[17] Park Ch.W., Seo D.R. Sentiment analysis of Twitter corpus related to artificial intelligence assistants // 2018 5th.

[18] Galinsky R., Alekseev A., Nikolenko S.I. Improving neural network models for natural language processing in russian with synonyms // 2016 IEEE Artificial Intelligence and Natural Language Conference (AINL).

[19] Rubtsova Y.V., A method for development and analysis of short text corpus for the review classification task, Proc. Trudy XV Vserossiiskoy Naychnoy Konferencii RCDL, pp. 269-275, 2013.

[20] Go A., Bhayani R., Huang L. Twitter sentiment classification using distant supervision //CS224N project report, Stanford. – 2009. – Т. 1. – №. 12. – С. 2009.

[21] Hussein, Sherif, Twitter Sentiments Dataset, Mendeley Data, V1, 2021, doi: 10.17632/z9zw7nt5h2.1