2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA) | 978-1-6654-3981-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/SUMMA53307.2021.9632004

Studying Comments on Russian Patriotic Actions: Sentiment Analysis Using NLP Techniques and ML Approaches

Anton Sysoev
Department of Applied Mathematics
Lipetsk State Technical University
Lipetsk, Russia
sysoev_as@stu.lipetsk.ru

Andrei Linchenko
Philosophy Department
Lipetsk State Technical University
Lipetsk, Russia
linchenko@mail.ru

Vladimir Kalitvin
Department of Mathematics and Physics
Lipetsk State Pedagogical University named after P.P. Semenov-Tyan-Shansky
Lipetsk, Russia
kalitvin@gmail.com

Daniil Anikin
Department of Political History
Lomonosov Moscow State University
Moscow, Russia
dandee@list.ru

Oksana Golovashina
Laboratory for Comparative Studies of Tolerance and Recognition
Ural Federal University
Yekaterinburg, Russia
ovgolovashina@mail.ru

***Abstract*—Due to the increasing number of online social media resources and social networks the problem of natural language processing (NLP) applying machine learning (ML) approaches is becoming an important concept in monitoring non-formalised text data and catching information in real time. Studies in historical memory and appropriate patriotic actions have to be based on detailed people's moods examination. The presented study contains approaches to analyse the polarity of comments on patriotic actions in Russia. Described scheme of NLP text processing and ML classifier were used in the analysis of two regions. There were used standard approaches to convert natural texts into numerical representation. To identify the polarity of the text several models including linear and deep neural networks were examined. Long-short-term-memory network has demonstrated the best result on validation set and was used to the solve the stated problem.**

***Index Terms*—Sentiment Analysis, Natural language processing, Machine Learning algorithms, Analysing comments, Patriotic Actions**

I. Introduction

In recent years social media sources, forums, blogs, and other forms of online communication tools have radically affected everyday life, especially how people express their opinions. At the same time comments can present the real attitude of people to meaningful social problems or inner policy strategies. In this regard, the study of topical problems of modern Russian society in the light of online communications is becoming increasingly important. One of these problems, which have become the subject of extensive discussions in recent years, is the problem of the population's attitude to the commemoration of the Victory Day. The relevance of these discussions is associated primarily with the fact that in the conditions of the ideological vacuum of the 1990s - 2000s. it is the memory of the Victory in the Great Patriotic War of 1941-1945 continued to be one of the most important factors in the spiritual cohesion of the population of Russia, as well as the most important symbolic resource of the state politics of memory. The year 2020 in the Russian Federation was officially named "The Year of Memory and Glory", which was associated with the celebration of the 75th anniversary of Victory in the Great Patriotic War of 1941-1945. The symbolic significance of this date for modern Russian society does not mean that there is no definite dynamics in the perception of war and its individual images in the historical consciousness. It could be said that the memory of the war demonstrates a tendency towards fragmentation and transformation within the framework of individual regional strategies of referring to the past, and it is the Internet environment, where the expression of opinions is not limited to group and institutional values, demonstrates special variability. In this regard, the team of authors focused on a comparative analysis of the comments of visitors to news sites in the Tomsk and Tula regions, dedicated to commemorations of Victory Day in the period from 2015 to 2021. The choice of these areas was associated with a working

The reported study was funded by RFBR and EISR according to the research project № 21-011-31153

978-1-6654-3981-7/21/$31.00 ©2021 IEEE

hypothesis, according to which the state politics of memory in relation to the Great Patriotic War was preceded by the existence of heterogeneous cultures of memory, which varied significantly depending on the region. The most significant factor that influenced the formation of these cultures of memory was the direct impact of the region by the war, therefore, as M. Gabovich's studies demonstrate, the formation of the all-Union image of the Great Patriotic War in the mid-1960s was based on the existing local commemorative practices in the western and central regions of Soviet Union. From this point of view, it seems logical to compare two regions with different historical backgrounds: the Tula region, on the territory of which the hostilities took place, and the Tomsk region, which was deep in the rear.

These regions are comparable in terms of population size, gender and age composition, as well as population distribution by rural and urban settlements. The choice of the chronological period was associated with the general updating of the state politics of memory and the increased attention of the authorities to the commemorations of Victory Day after 2014. This period also saw the flourishing of the social and patriotic action ”Immortal Regiment”, which became an important commemorative practice not only of official events, but also of the family memory of Russians. The main goal of our article was to study the dynamics of statements on the news Internet sites of the Tomsk and Tula regions about the celebration of Victory Day and its commemorative actions. However, the work with a large number of comments required us to turn to the use of artificial intelligence methods for processing Big data, which acquired the character of an independent task of this study. Accordingly, in this article, we will not only present the results of processing and analysis of data from Internet comments regarding the celebration of Victory Day in two regions (Tomsk and Tula regions), but also the results of a study of the specifics of the application of artificial intelligence methods in relation to the problem of interest to us.

The problem stated below can be automatically solved by applying Sentiment Analysis approaches containing tasks of emotions detection, their classification, and the mining of opinion polarity and subjectivity [1]. The analysis can be performed at different levels of granularity, ranging from the sentiment of the whole document or each sentence to an opinion of a certain entity or aspect. There are two well-known and widely-used category of approaches to Sentiment Analysis: (1) lexicon-based approaches using lexicon (a vocabulary, a list of words, a dictionary associated with their polarities) to find a polarity to a new text and (2) Machine Learning (ML) approaches involving large manually labeled data sets. Many algorithms can be used to gain labels to a new text starting from classical methods (line logistic regression, decision trees, etc.) to deep neural network models.

II. State-of-Art in Sentiment Analysis of Russian Texts

Currently there are only a couple case studies in Sentiment Analysis applied to the Russian-language short texts like tweets or comments in social networks. Among them is the study [2] focusing on researching short texts from Social networking sites such as Facebook, Twitter and VKontakte, online stores such as eBay, Amazon and Alibaba. They have compiled and summarised approaches and methods for text acquisition, text preprocessing and sentiment classification used by various authors applied to the Latvian and Russian languages. The paper [3] contains practical recommendations to help scholars selecting an appropriate training data set as well as performs an additional literature review on publicly available sentiment data sets of Russian-language texts. The list of data sets as well as their detailed explanation is presented in study [4] containing also the review of different methods for obtaining word and sentence embeddings [5], [6], ranging from simple Bag-of-Words [7] and Word2Vec [8] models to pre-trained language models as ELMo or BERT.

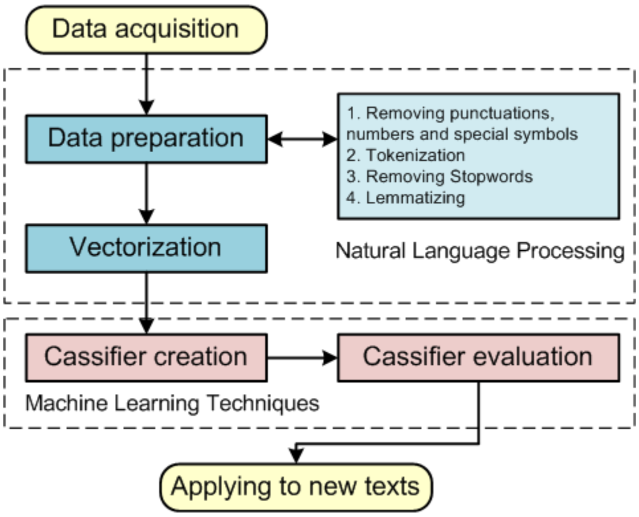

Fig. 1. Baseline of the study

The presented study uses a traditional baseline involving data acquisition, data preparation, classifier creation, classifier evaluation and finally its implementation to classify unseen texts (cf. Figure 1). To prepare data for model training and further evaluation procedures NLP approaches were used, as well as there were applied ML techniques to build the model and to classify new texts. The target group of comments are those related to news on Russian patriotic actions (Victory Day anniversaries, Immortal Regiment, commemorations of 22th June) conducted within several years in social media resources and social networks in Tomsk and Tula regions.

III. Train Data Set Description and Test Data Acquisition

The applied problem required a specific training data set choosing. Many of existing Russian-language data sets are reviews on goods labeled as positive and negative or news. Since the goal of the study is to identify the longitudal effect of organized events on peoples' opinion and to find possible regions with the growing peoples' dissatisfaction, it

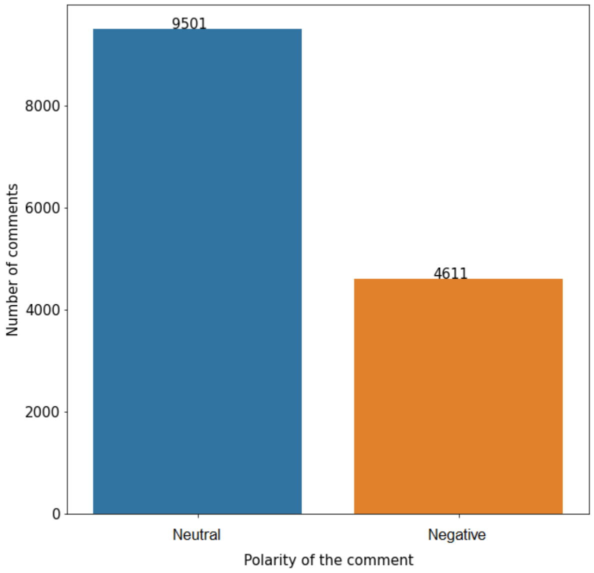

Fig. 2. Structure of train data set

was supposed to label comments with two available classes — neutral or negative. The best results were obtained by using Russian Language Toxic Comments data set available at Kaggle platform [9]. These data are formed by 14 112 comments of Russian spoken language phrases ans sentences including taboo words. The proportion of neutral and negatives realisations in train data is 67% of neural and 33% of negative comments (cg. Figure 2). To check further models and choose the best one it was formed balanced test data set which was not seen by the models before.

Firstly by sociologists and historical memory researchers there were selected open resources with comments on specified topics. To parse the data there was developed a program code in the Python programming language automatically connecting to a resource from the list, catching data and converting it into a csv-format file. In this case two approaches were used. In the first case, to receive an html document with comments, it was enough to execute a request at a given address, analyze the structure of the document, and extract the data. For pages containing comments loaded using fragments of javascript code, a version of the parser was developed that emulates the operation of a browser, which made it possible to get all comments from the specified addresses. When developing the program code, the lxml, bs4, selenium libraries were used.

IV. Models to Classify Comments

A. Random Forest

The random forest being an “ensemble learning” technique consists of the aggregation of a large number of decision trees, resulting in a reduction of variance compared to the single decision trees [10]. Each tree of the forest is built based on a bootstrap sample drawn randomly from the original data set using the CART method and the Decrease Gini Impuritiy as the splitting criterion. Random forest is usually considered a black-box algorithm, as gaining insight on a prediction rule is hard due to the large number of trees.

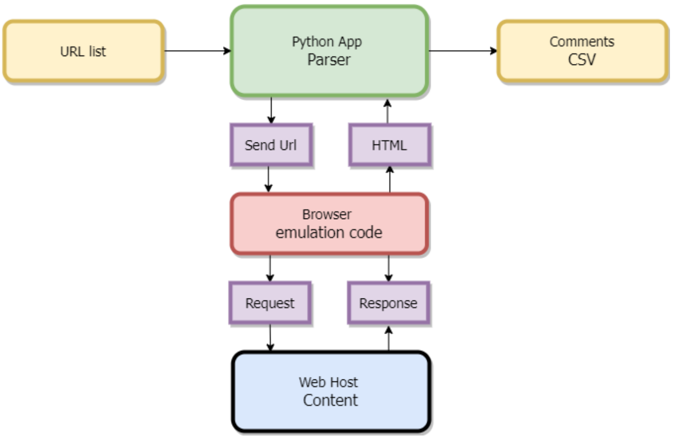

Fig. 3. Data collection diagram

B. Logistic Regression

Let Y is the binary output and $x_1, ..., x_n$ are random input variables, termed features in the Sentiment Analysis. The logistic regression links the conditional probability $P(Y = 1|x_1, ..., x_n)$ to $x_1, ..., x_n$ through the following model

$$P(Y = 1|x_1, ..., x_n) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}, \qquad (1)$$

where $\beta_0, \beta_1, ..., \beta_n$ are regression coefficients estimated by maximum-likelihood method from the considered data set.

Applying (1) the new realization is assigned as

$$Y = \begin{cases} 1, & \text{if } P(Y = 1) > \alpha, \\ 0, & \text{otherwise.} \end{cases} \qquad (2)$$

The commonly used threshold in the rule (2) $\alpha = 0.5$, a so-called Bayes classifier.

C. Convolutional Neural Network

A convolutional neural network (CNN) is usually a combination of convolutional layers, subsampling layers, and if there are fully-connected layers on the output. All three types of layers can alternate in any order [11].

Within the convolutional layer, neurons that use the same weights are combined into feature maps, and each neuron of the feature map is connected to a part of neurons from the previous layer. When calculating the network, it turns out that each neuron performs the convolution of some area of the previous layer (defined by the set of of neurons connected to this neuron). The convolutional layer is similar to the application of the convolution operation, where only a matrix of small (convolutional kernel), which is “moved” throughout the processed layer. Another feature of the convolutional layer is that it slightly reduces the image due to edge effects. Subsampling layers perform dimensionality reduction (usually by several times). This can be done in different ways, but often the max-pooling method is used, i.e. the whole feature map is divided into cells, from which the maximal ones are selected.

D. Reccurent Neural Network. LSTM

Neural networks make predictions independently on each object in the sample. In some problems, however, the order of the objects is important. In general, recurrent neural networks (RNN) [12] can take a sequence or a single object as input and have a sequence of outputs or a single output. All RNN have a chain structure — repeating cell. These units sequentially process word vectors according to the following scheme

$$h_t = f(U_t \cdot x_t + W \cdot h_{t-1} + b), \tag{3}$$

where h_t is the cell output, f is the activation function, weights U_t are individual for each element, weights W are constant.

LSTM (long-short-term-memory) are models capable of finding long- and short-term dependencies. Each cell consists of four connected neurons. A unit has two recurrent components: output vector h_t and state vector c_t. The distinguishing feature is that LSTM does not use activation functions inside the component c_t. It passes directly through the entire chain, participating only in a few linear transformations. Thus, the stored value is not blurred in time, and the gradient or penalty does not disappear when the error propagation backward in time is used in training the network. LSTM can remove information from recurrence components. This process is governed by filters. The first step in LSTM is to determine what information can be discarded from the vector of states. For this purpose exists a neuron with a sigmoid activation function. This layer takes a vector (h_{t-1}, x_t) as input, and returns a vector $f_t \in [0, 1]^m$:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f).$$

The next step is to decide what new information will be stored in the state vector. This step consists of two parts. First, the sigmoid layer determines which components of the state vector and to what extent should be updated. Then the next layer builds a vector of new values with which to replace the vector components that need updating. The vector (h_{t-1}, x_t) is input, the outputs are vectors of dimension m:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{c_t} &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c). \end{aligned}$$

New vector of states:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c_t}.$$

It is necessary to update the output vector. For this purpose several filters are applied to the vector of states. First is the sigmoid layer, then $\tanh-$layer. The output values from the range $[-1; 1]$ are multiplied with the output values of the sigmoid layer, which makes it possible to output only the required information.

E. Reccurent Neural Network. GRU

GRU (gated recurrent units) are a gating mechanism in RNN similar to LSTM with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. The state vector is not used in this modification, its role is taken by the output vector, which passes through the circuit directly and is not subject to the activation function. The unit consists of three connected neurons. First, a vector (h_{t-1}, x_t) is fed to the input of the sigmoid layer, at the output we get the update z_t, whose dimension coincides with the vector h_{t-1}. Then using another sigmoid layer similarly a vector r_t is constructed, specifying which components of the vector h_{t-1} should be used to build the output vector update in what degree. At the third neuron with the activation function $\tanh$ a vector composed of the subordinate product r_t, h_{t-1}, x_t is fed. This updates the output vector:

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]), \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]), \\ \tilde{h_t} &= \tanh(W \cdot [r_t \cdot h_{t-1}, x_t]), \\ h_t &= (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h_t}. \end{aligned}$$

F. Comparing Results

All model described above were realized with Python functions with tensorflow and keras libraries. To choose the most effective model there were analysed following quality metrics:

- Accuracy being the ratio of true positives and true negatives to all positive and negative observations;
- Precision representing the model's ability to correctly predict the positives out of all the positive predictions it made;
- Recall representing the model's ability to correctly predict the positives out of actual positives, and
- $F-$score giving equal weight to both the Precision and Recall for measuring its performance in terms of accuracy, making it an alternative to Accuracy metrics:

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

TABLE I
COMPARATIVE ANALYSIS OF THE STUDIED MODELS

	Accuracy	Precision	Recall	F-score
Logistic regression	0.74	0.59	0.85	0.58
Random forest	0.79	0.70	0.83	0.72
CNN	0.89	0.85	0.85	0.85
LSTM	0.88	0.88	0.88	0.88
GRU	0.87	0.85	0.84	0.84

Table I presents the comparing results of analysed models. Expectedly the worst results were shown by the logistic regression model. Compared to the random forest model the neural network models deliver higher accuracy of text polarity recognition. The highest accuracy among neural network procedures is demonstrated by LSTM networks. However, during their training more parameters are used, which noticeably increases the speed of operations. Based on these results the LSTM model pre-trained on the mentioned data set was used to solve the applied problem.

V. Comments on Russian Patriotic Actions: Main Findings

As it was mentioned above there were explored several open sources with free comments options. To perform Sentiment Analysis with the chosen LSTM model there were retrieved 3674 raw comments within the last 7 years from 2015 to 2021. Such longitudinal analysis can help in studying negative attitude formation in the chosen Russian regions. Radar charts in Figure 4 show the distribution of comments throughout years.

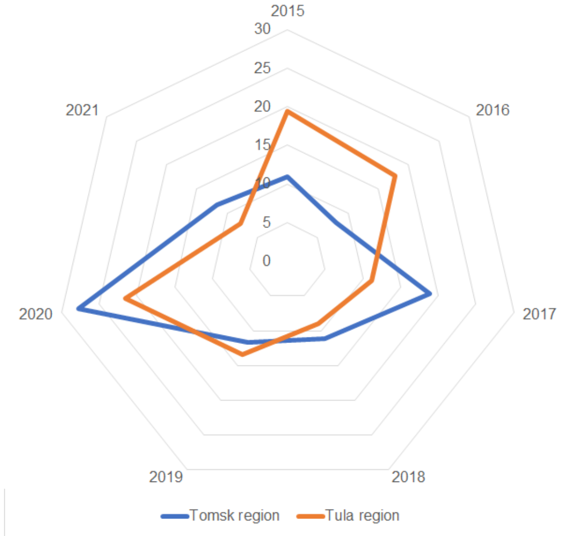

Fig. 4. Percentage distribution of comments in Russian regions throughout years

Figure 5 establishes the dynamics of negative comments in Tomsk and Tula regions.

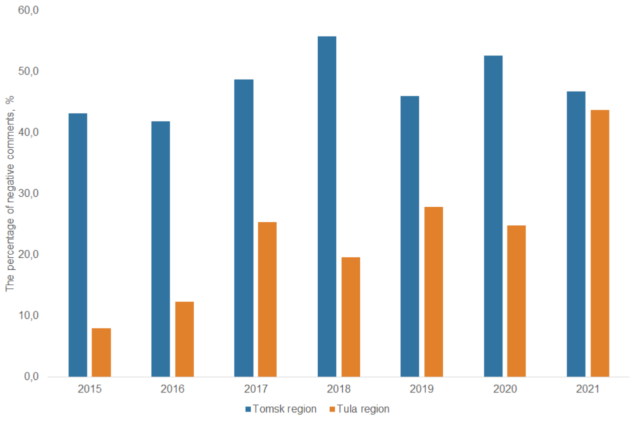

Fig. 5. Dynamics of negative comments in Russian regions throughout years

The presented diagrams allow us to draw primary conclusions about the regional memorial landscape of modern Russian society. Diagram on Figure 4 demonstrates a stable pattern of updating the memory of the Great Patriotic War in accordance with the state commemorations associated with the anniversary dates of the end of the war. In particular, the 75th anniversary of the Victory in 2020, actively covered through state information channels, caused a pronounced surge of interest in both the Tula and Tomsk regions. While the previous anniversary of the Victory in 2015, considered as "another" and not very significant from the official point of view, led to the preservation of the average parameters of interest in this event.

In our opinion, the analysis of the dynamics of negative comments in relation to the Great Patriotic War on information portals and in social networks is very indicative (diagram on Figure 5). Attention was drawn to the fact that the number of negative comments does not directly correlate with anniversary / non-anniversary years, at least no ups or downs in 2015 and 2020 are noted. In the Tomsk Region, the number of negative comments in general continues to remain stable throughout the entire period studied. While in the Tula region, compared to 2015, the number of such comments is growing steadily, and not in an explosive manner, but consistently, which indicates a certain trend in relation to the Great Patriotic War that needs to be understood.

In general, during the studied period, one can see a certain difference in the local cultures of memory. So, the memorial culture of the Tula region initially (2015) turns out to be much more loyal to the memory of the war, but over the studied period there is a certain leveling, which manifests itself in an increase in the number of negative comments and the achievement of quantitative values of the Tomsk region in this parameter.

VI. Conclusion

The initial results obtained by us in the course of the research indicate the enormous methodological potential of using artificial intelligence and machine learning methods in studying the problem of perceiving commemorations of the events of the Great Patriotic War of 1941-1945 and Victory Day in the Russian Internet space. The results of the study allow us to confirm the working hypothesis, according to which the state politics of memory in relation to the Great Patriotic War should take into account the existence of heterogeneous cultures of memory of the war, which vary significantly depending on the region. Our results convincingly demonstrate a stable pattern of updating the memory of the Great Patriotic War in accordance with state commemorations associated with the anniversary dates of the end of the war. It was found that the number of negative comments does not directly correlate with anniversary / non-anniversary years. At the same time, there was a tendency towards an increase in negative comments, which was explained not so much by the disappointment of the population in the commemorations themselves, but by the deterioration of the current social and economic agenda, the growth of social apathy and conflict in Russian society.

It should be noted that there are language transformer models like ELMo, BERT or Dostoevsky model pre-trained on Russian language corpus. The prospective step is applying these models to the comments on Russian patriotic actions or the other social valuable problems analysis.

The results of the study convincingly show the need to expand the sample of regions and further differentiate the studied Internet sources. Regarding the topic of perception of Victory Day in the Russian Internet space, it should be noted that a purely quantitative analysis of comments on news sites and social networks has only a partial explanatory value. It allows you to capture the emotional richness of messages, but does not allow you to explain the reason for this richness. For example, the statement of the growth of negative comments in itself does not allow us to identify the reasons for this growth, since hypothetically, the reason for the increase in such assessments may be not only and not so much a change in attitudes towards the Great Patriotic War itself, but rather specific commemorative practices of regional authorities, or the general level of growth of opposition sentiments by region. That is why the quantitative analysis of Big data should be accompanied, in our opinion, by a discourse analysis of the information received, which will significantly correct and explain those patterns that are revealed through machine analysis.

References

[1] H. Soong, N. B. A. Jalil, R. K. Ayyasamy, and R. Akbar, "The Essential of Sentiment Analysis and Opinion Mining in Social Media," 2019 IEEE 9th Symp. Comput. Appl. Ind. Electron., 2019.

[2] R. Vīksna and G. Jēkabsons, "Sentiment Analysis in Latvian and Russian: A Survey," Appl. Comput. Syst., 2018.

[3] S. Smetanin, "The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives," IEEE Access, 2020.

[4] S. Smetanin and M. Komarov, "Deep transfer learning baselines for sentiment analysis in Russian," Inf. Process. Manag., 2021.

[5] J. Rodina, D. Bakshandaeva, V. Fomin, A. Kutuzov, S. Touileb, and E. Velldal, "Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: the Case for English, Norwegian, and Russian," 2019.

[6] V. Malykh, A. Alekseev, E. Tutubalina, I. Shenbin, and S. Nikolenko, "Wear the right head: Comparing strategies for encoding sentences for aspect extraction," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019.

[7] S. Smetanin and M. Komarov, "Sentiment analysis of product reviews in Russian using convolutional neural networks," in Proceedings - 21st IEEE Conference on Business Informatics, CBI 2019, 2019.

[8] A. Alekseev and S. Nikolenko, "User profiling in text-based recommender systems based on distributed word representations," in Communications in Computer and Information Science, 2017.

[9] A. Belchikov, Russian language toxic comments. Available at https://www.kaggle.com/blackmoon/russian-language-toxic-comments

[10] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," BMC Bioinformatics, 2018.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, 1998.

[12] P. Liu, X. Qiu, and H. Xuanjing, "Recurrent neural network for text classification with multi-task learning," in IJCAI International Joint Conference on Artificial Intelligence, 2016.