data and the random forest method is used to fill in missing data values. Then you use the gradient-boosting decision tree method to view important indicators, process proven indicators using a metric system model based on logistic regression, and get a personal credit score. Finally, the model is tested using a BP neural network, and the model is used to predict the level of personal credit. The study shows that machine learning can further improve the accuracy of individuals' credit ratings and provide a scientific basis and background information for commercial banks' credit ratings.

***Key words**: big data, credit scoring, logistic regression, machine learning, data desensitization, Decision tree, BP neural network.*

Сведения об авторах

Жанна Муратовна Ордабаева – PhD докторант кафедры «Программная инженерия»; Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

Айман Николаевна Молдагулова – кандидат физико-математически наук кафедры «Физика и математика»; профессор кафедры «Программная инженерия»; Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

Авторлар туралы мәліметтер

Жанна Муратовна Ордабаева – «Бағдарламалық қамтамасыз ету» кафедрасының PhD докторанты; Қ.И.Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

Айман Николаевна Молдагулова – физика-математика кафедрасының физика-математика ғылымдарының кандидаты; Бағдарламалық қамтамасыз ету инженериясы кафедрасының профессоры; Қ.И. Сәтбаев атындағы Қазақ ұлттық зерттеу техникалық университеті, Алматы, Қазақстан; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

Information about authors

Zhanna Ordabayeva – PhD doctoral student of the department "Software Engineering"; Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan; e-mail: zhannaordabayeva@gmail.com. ORCID: 0009-0009-8273-3147.

Ayman Moldagulova – Candidate of Physical and Mathematical Sciences of the Department of Physics and Mathematics; Professor of the Department of Software Engineering; Kazakh National Research Technical University named after K. I. Satpayev, Almaty, Kazakhstan; e-mail: a.moldagulova@satbayev.university. ORCID: 0000-0002-1596-561X.

Материал поступил в редакцию 01.03.2023 г.

DOI: 10.53360/2788-7995-2023-1(9)-8

МРНТИ: 20.19.27

K. Nursakitov*, A. Bekishev, S. Kumargazhanova, A. Urkumbaeva
D. Serikbayev East Kazakhstan technical university,
070004, The Republic of Kazakhstan, Ust-Kamenogorsk, 69 Protozanov Street
*e-mail: nursakitov@bk.ru

REVIEW OF METHODS FOR DETERMINING THE TONATION OF TEXTS IN NATURAL LANGUAGES

***Annotation:** The analysis of sentiment in user comments finds application in many areas, such as evaluating the quality of goods and services, analyzing emotions in messages, and detecting phishing advertisements. There are numerous methods for analyzing the sentiment of textual data in the Russian language, but automatic sentiment analysis of Russian-language texts is much less developed than for other major world languages. This article is part of a broader study on the creation*

of an information system for detecting dangerous content in the cyberspace of Kazakhstan. The purpose of this article is to provide an analytical review of the different approaches to sentiment analysis of Russian-language texts and to compare modern methods for solving the problem of text classification. Additionally, the article seeks to identify development trends in this area and select the best algorithms for use in further research. The review covers different methods for text data preprocessing, vectorization, and machine classification for sentiment analysis of texts, and it concludes with an analysis of existing databases on this topic. The article identifies some of the main unresolved problems in sentiment analysis of Russian-language texts and discusses planned further research.

Key words: *machine learning, recurrent neural networks, natural language processing, text sentiment, NLP, information Technology.*

Introduction

Text sentiment analysis is the process of automatically determining the emotional coloring of the text, expressed in relation to some object, topic or event. The main task of sentiment analysis is to determine whether the sentiment of the text is positive, negative or neutral[1]. Sentiment analysis can be performed using a variety of methods, including rules, machine learning, and deep learning. Some methods use a dictionary containing a list of positive and negative words, as well as their weights. When analyzing a text, words are searched in the dictionary and their weight coefficients are added up, which makes it possible to determine the tone of the text. Sentiment analysis can be used to automatically process reviews, measure public opinion, evaluate brand reputation, and detect signs of fraud in ads or social media posts..

Sentiment analysis of texts occurs in several stages (Fig. 1). At the first stage, the source text is preprocessed, then informative features are extracted (text vectorization), a sentiment classifier (recognizer) is built on their basis, and the last stage is the evaluation of the result of the work. The stage of text vectorization for linguistic classification methods is not mandatory, since such classifiers work directly with texts, and not with their vectors.

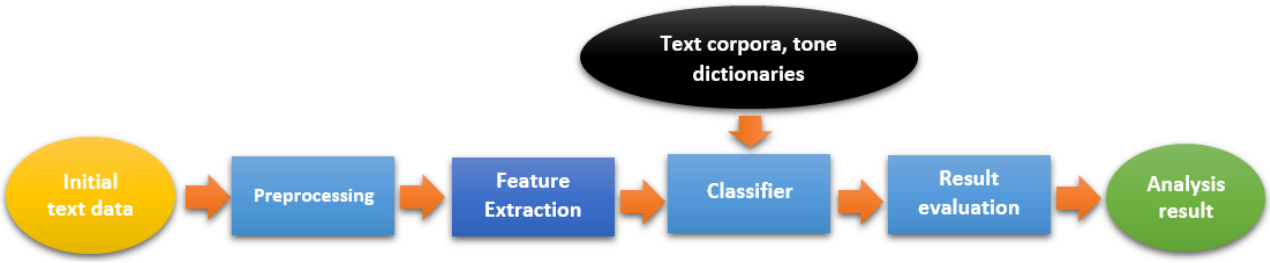

Figure 1 – The stages of sentiment analysis of text

Text preprocessing

Text preprocessing is the first step in its analysis. It is necessary in order to extract relevant information from the "noisy" text. Text preprocessing includes converting all words to a single case, removing punctuation marks, removing stop words, tokenization, word normalization, and, if necessary, other operations.

When converting all words to the same case, as a rule, all uppercase characters are converted to their lowercase forms, since it is assumed that the uppercase or lowercase forms of words do not differ. All texts contain punctuation marks, which most often perform a syntactic function, so when analyzing emotions in a text, there is no need to save them. Also, when processing the text, stopwords are removed – words that do not contain a semantic load, for example, prepositions, conjunctions, particles, etc. A necessary preprocessing step for subsequent computer analysis of the text is tokenization of words – splitting the text into separate meaningful units (tokens) [2]. The easiest way to tokenize Russian-language text is to split it into words by spaces. Word paradigms in Russian have a large number of word forms that convey the same meaning. The form of a word does not always carry useful information, therefore, when analyzing a text, it is recommended to normalize all words, i.e., represent the word in its initial form. Normalization can be done in two ways: lemmatization and stemming. Lemmatization is the transformation of a word to its initial form (lemma). Lemmatization is based on a morphological dictionary. If the word is not present in the dictionary, then a hypothesis is constructed about ways to change the word and obtain a lemma for it. Stemming – obtaining the basis of a word, while endings, suffixes, and prefixes are

discarded from words. Thus, all words in the text are reduced to a single form. Stemming is based on morphological rules and does not require a dictionary.
Each of the stages of text preprocessing allows you to reduce the size of the space. Depending on the source text, preprocessing may include only a few operations, and each operation may be refined manually, taking into account all exceptions.

Extracting features from text

Before using a machine classifier, it is necessary to present the text in a numerical form (featured description), i.e. vectorize the text. Let's consider several modern ways of text vectorization.

BoW (Bag of Words) is one of the simplest and most common text vectorization methods. It is based on the fact that the vector of each document in the feature space is formed from the frequency of occurrence of individual words in this document [3]. This method completely ignores the word order in the text and takes into account only their presence.

One-hot encoding (direct encoding) is a vectorization technique that is used to convert categorical (nominal) data into numeric vectors [4]. It is based on the creation of fixed-length vectors, where each element of the vector corresponds to one of the possible values of a categorical feature.

SVD (Singular Value Decomposition) – A text vectorization method that is based on the assumption that words that often occur together have a stronger relationship than words that rarely or never occur together[5]. It allows you to represent the word-document matrix as a product of three matrices of lower rank.

Word2Vec (a toolkit developed by Google) is a text vectorization method that allows words to be represented as vectors of numbers corresponding to their semantic meaning [6]. This method is used to analyze text data and is widely used in machine learning tasks such as text classification, searching for similar documents, and text generation.

GloVe (Global Vectors for Word Representation) is a text vectorization method that allows you to represent words as vectors based on the statistical properties of their interaction in texts. GloVe is based on the idea that semantically related words often appear in the context of each other. However, unlike Word2Vec, GloVe uses a co-occurrence matrix to determine the proximity between words [8].

BERT (Bidirectional Encoder Representations from Transformers) is a text vectorization method based on Transformers and trained on a large amount of text. [9]. BERT uses a layered architectural approach and a bidirectional encoding model, which allows it to take into account the context and dependencies between words in a sentence when generating vector representations. Unlike other models that only look at sentences in one direction (from front to back or vice versa), BERT analyzes a sentence from both sides [10]. This is a lighter and faster version of BERT that roughly matches its performance. The authors of [11] showed that transferring training from a multilingual BERT model to a monolingual model for the Russian language leads to a significant increase in performance when analyzing emotions in a text.

ELMo (Embeddings from Language Models) is a text vectorization method that uses deep language models such as LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks) to create word embeddings [12]. Unlike other vectorization methods, ELMo builds word embeddings based on the context in which they are located, taking into account both the left and right contexts. This allows you to create more accurate embeddings that take into account not only the word itself, but also its context, which is especially useful in the case of synonyms or words with multiple meanings.

Sentiment classification of text data

To date, there are a large number of methods for determining the tone of the text [13]. All of them can be divided into three main groups (Fig. 2).

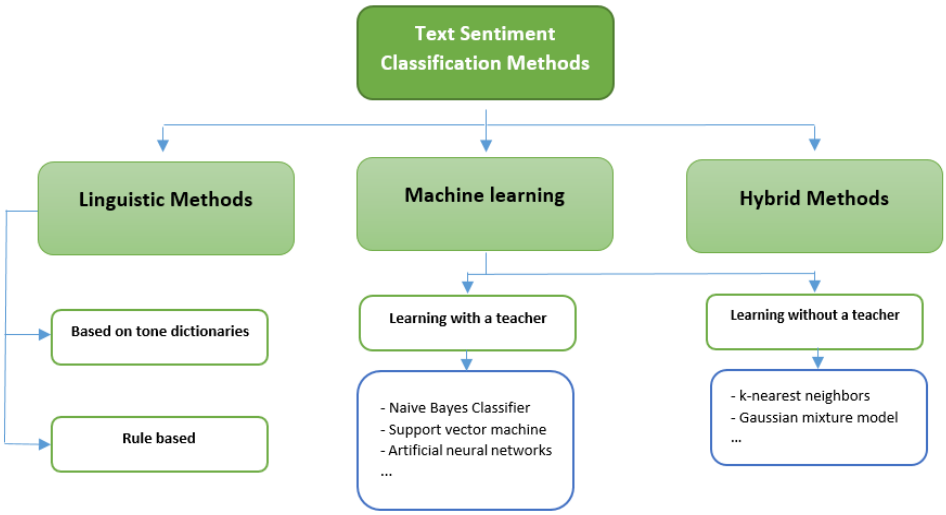

Figure 2 – Text sentiment. Classification mehods

The essence of the first linguistic method based on tone dictionaries is that for each word from the text its tone is determined (for example, positive, negative or neutral) based on tone dictionaries. These dictionaries contain a list of words that refer to positive or negative sentiment. Then, for each word in the text, its weight is determined, corresponding to its tonality. For example, a positive word can have a weight of 1, a negative word – 1, and a neutral word 0. Then all the weights of the words in the text are summed up, and based on this, the overall sentiment of the text is calculated. This method was used for sentiment analysis in [14].

The second linguistic method is based on rules. For this method to work, a large set of production rules for the “if –> then” construction is needed. This method also involves the use of tone dictionaries in which the words belong to a particular class. The problem of sentiment analysis is solved using a rule-based method, for example, in [15].

Machine learning methods can be divided into two main categories: supervised learning and unsupervised learning. Supervised learning is a machine learning technique in which a model is trained on data that contains correct answers, i.e. data is marked. In this case, the model finds dependencies between inputs and outputs in order to later predict responses to new data [16]. Examples of supervised learning algorithms: linear regression, logistic regression, decision trees, random forest, gradient boosting, and neural networks [17].

Unsupervised learning is a machine learning technique in which a model is trained on unlabeled data with no explicit answers. In this case, the model itself finds patterns in the data, groups data by similarity, and finds hidden dependencies between features. [18] Examples of unsupervised learning algorithms: clustering, principal component analysis, association rules, density-based learning algorithms, and autoencoders [19].

There are also hybrid methods that combine several different methods. In [20], for the problem of text classification, a hybrid method was used that combines tone dictionaries and the support vector machine. In [21], the authors combined CNN and k-nearest neighbors to solve the sentiment analysis problem.

After the classification stage of sentiment analysis of texts, a quantitative assessment of the results follows, which can be carried out using a set of the following statistical indicators: accuracy or precision, completeness (recall) and F-measures (F-score) [13].

Corpora for Sentiment Analysis of Texts

Despite the relevance of the sentiment analysis of Russian-language texts, the number of annotated corpora for the Russian language is small. At the beginning of 2022, we managed to find four tone dictionaries and seven text corpora in the public domain, designed for the task of sentiment analysis of Russian-language texts.

Russian-language tonal dictionaries in the public domain

When using a method based on tone dictionaries, for automatic text classification, it is necessary to rely on a dictionary that contains words with markings for belonging to a certain sentiment. Markup can be binary (2 classes), ternary (3 classes) and multi-class (more than three classes). There are several tonal dictionaries for the Russian language.

WordmapSent [22] is a tonal dictionary covering more than 46 thousand words of the Russian language. In the published dataset, each input is associated with a tonality label and a numerical value of the strength of the emotional-evaluative charge from a continuous range [1].

The RuSentiLex tonal dictionary [23] can contain both individual words and phrases, for which their characteristics are indicated, denoting the part of speech or the syntactic type of the group, their lemmatized form, tonality, and source of information. Depending on the context, the same word can take on a different meaning. Therefore, the authors of the dictionary introduced a separate class of sentiment, denoting a mixed assessment of the word. Also, the authors partly solved the problem with words that have several meanings. They list all the meanings of the word according to the RuThes thesaurus [24] and give a reference to the corresponding concept, the name of the concept is written in quotation marks. In such cases, each meaning of the word is assigned its own sentiment value.

LinisCrowd [25] is a tone dictionary based on user-generated Internet content on social and political topics. Initially, the dictionary was compiled from marked-up texts obtained from the social network Facebook. Subsequently, the dictionary was expanded by adding other word forms to it, as well as words from other dictionaries.

WordNetAffect [26] is a lexical resource that contains words that describe emotions. It was created on the basis of the ontology of WordNet - the semantic lexicon of the English language - by selecting and labeling sets of synonyms (synsets) with emotional concepts. The sets of synonyms were manually labeled with emotional labels, then they were additionally re-labeled into six emotional categories. For the Russian language, the authors of the dictionary manually translated WordNetAffect synsets from English.

The tonal dictionary from Belyakov's work [27] contains 690 bases of emotional words. The dictionary is divided into two classes: the basics of Russian words with positive and negative emotional coloring.

Russian-language emotionally colored text corpora in the public domain

There are several emotionally colored text corpora for the Russian language, their brief description is presented below.

The largest Russian conference on computational linguistics "Dialogue" annually holds competitions in computer analysis of the Russian language (http://www.dialog-21.ru/evaluation/), one of the main areas of competition is the analysis of the tone of texts. Yes, in 2015 and 2016. the organizers provided SentiRuEval text corpora. SentiRuEval-2015 [28] contains reviews collected from the Twitter network about restaurants and cars. In addition to the general tone of the review, SentiRuEval-2015 contains various target aspects of the object being evaluated. Each of these aspects can also have a tonal value. SentiRuEval-2016 [29] includes reviews about banks and mobile operators collected from Twitter. Feedback markup shows the object of the feedback and the relationship of the subject to this object.

LinisCrowd [25] is a collection of documents on socio-political topics. The records of the blogging platform "LiveJournal" were used as a data source. RuSentiment [22] is a text corpus that includes posts collected from the VKontakte social network on various topics. Some posts may not be marked by tone, but they may belong to a certain class of utterance (template greetings, thank you and congratulations messages). RuTweetCorp [30] is a corpus of Russian-language twitter posts automatically categorized into two classes. The ROMIP 2012 [31] and Auto_reviews [32] corpora are also freely available.

Sentiment analysis can also be used in the analysis of conversational speech of speakers. To solve this problem, you can use the multimodal RAMAS corpus [33]. It contains about seven hours of audio and video recordings of interactive dialogues enacted by several actors. Before analyzing the text component of the speakers' statements, you first need to get the spelling transcription of the audio files, which is not provided by the developers.

Software products for sentiment analysis of Russian texts

The task of determining the tone of a text is commercially in demand; therefore, various oriented computer systems are being developed that analyze the tone of texts. At the beginning of 2020, we were able to find five freely available software systems designed for sentiment analysis of Russian-language texts.

SentiFinder [34] is a software module of the high-speed linguistic text analysis system Eureka Engine. It determines the tone of texts in Russian, English and Armenian. A feature of this module

is that it allows you to assess the degree of emotionality of the statement. It is designed to determine the tone of reviews for various products, as well as news feeds and blogs.

Semantria [35] is a sentiment analysis software module based on the Lexalytics platform. The system allows classifying the tone of messages in several European languages, including Russian. Semantria is designed for text analysis in the field of marketing.

SentiScan is a text sentiment recognition technology based on the YouScan platform [36]. The SentiScan classifier was trained on data that contained product reviews from various industries. YouScan is a commercial product, but it has a free trial available upon request.

SentiStrength is a software product for analyzing user sentiment [37]. It is designed to analyze short social internet texts. The result of the text analysis is two scores that take values from -5 (very negative) to 1 (not negative) and from 1 (not positive) to 5 (very positive). Initially, SentiStrength was developed for the analysis of the English language, but subsequently adapted for other languages, including Russian.

Texterra is an application for sentiment analysis of news messages [38]. The analyzed texts can be from specific areas: politics, finance, internet, health, and Twitter posts. The demo version of Texterra is freely available, its developers provide the ability to analyze the actual news collected from the Yandex.News platform and Twitter, as well as user texts entered manually.

As a rule, software products for sentiment analysis of texts in Russian are based on traditional teaching methods and do not use neural networks. This approach can be justified by the fact that neural networks require a large amount of training data, as well as a large amount of computational and time resources for their training.

Conclusion

The article presents an overview of approaches to sentiment analysis of Russian-language text data. The presence of numerous works on the topic of text sentiment analysis suggests that this task is relevant and commercially in demand in many areas, including advertising, politics, marketing, etc. This is confirmed by the increase in the number of conferences in the field of text analysis every year, as well as the number of publications on the analysis of both Russian-language data and texts in other languages. However, sentiment analysis systems for Russian-language texts are less developed than for the main world languages. Also, the Russian-language sentiment analysis shows a rather low accuracy compared to the English-language one, which is associated with the complex structure of the Russian language. To confirm this statement, one can consider works on the sentimental analysis of the Czech language, since the grammars of Russian and Czech are similar. The works [38-40] analyze the tonality of texts in English and Czech, and the results of the study show that the accuracy of sentiment recognition in English is higher than in Czech.

In further research, it is planned to implement an automatic classifier of advertisements for the presence of signs of fraud in them. To do this, it will be necessary to create an experimental corpus consisting of ads categorized as fraudulent and legal. Based on the obtained data, it will be possible to build a classifier, first using the method based on tone dictionaries, and subsequently other classification methods described in the article.

References

1. Enikolopov S.N., Kuznetsova Y.M., Smirnov I.V., Stankevich M.A., Chudova N.V. Creating a text analysis tool for socio-humanitarian research. Part 1. Methodical and methodological aspects. Artificial Intelligence and Decision Making. – 2019. – no. 2, pp. 28-38. doi:10.14357/ 20718594190203. (In Russian).

2. Polyakov E. V., Voskov L. S., Abramov P. S., Polyakov S. V. Generalized approach to sentiment analysis of short text messages in natural language processing. Informatsionno-upravliaiushchie sistemy [Information and Control Systems], 2020, no. 1, pp. 2–14. doi:10.31799/1684- 8853- 2020-1-2-14. (In Russian).

3. Soumya G. K., Joseph S. Text classification by augmenting bag of words (BOW) representation with co-occurrence feature. IOSR Journal of Computer Engineering, 2014, vol. 16(1), pp. 34-38.

4. Potdar K., Pardawala T. S., Pai C. D. A comparative study of categorical variable encoding techniques for neural network classifiers. International Journal of Computer Applications, 2017, vol. 175, no. 4, pp. 7-9.

5. Steinberger J., Jezek K. Text summarization and singular value decomposition Proceedings of International Conference on Advances in Information Systems, Springer, Berlin, Heidelberg, 2004, pp. 245-254.

6. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013. Available at: https://openreview.net/forum?id=idpCdOWtqXd60#7b076554-87ba-4e1e-b7cc-2ac107ce8e4d (accessed 2 May 2020).
7. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation. Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP-2014), 2014, pp. 1532-1543.
8. Pylieva H., Chernodub A., Grabar N., Hamon T. Improving automatic categorization of technical vs. Laymen medical words using fasttext word embeddings. Proceedings of the 1st International Workshop on Informatics and Data-Driven Medicine, IDDM 2018, 2018, pp. 93–102.
9. Devlin J., Chang M., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, vol. 1 (Long and Short Papers), pp. 4171-4186. doi:10.18653/v1/N19-1423
10. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (accessed 05 April 2020).
11. Kuratov Yu., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for russian language. Computational Linguistics and Intellectual Technologies, 2019, iss. 18, pp. 333-339 (In Russian).
12. Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations. NAACL-HLT. – 2018, vol. 1 (Long Papers). – pp. 2227-2237. doi:10.18653/v1/N18-1202
13. Dvoynikova A., Verkholyak O., Karpov A. Analytical review of methods for identifying emotions in text data. CEUR-WS, 2020. – vol. 2552. – pp. 8-21.
14. Tutubalina E.V., Ivanov V.V., Zagulova M.A., Mingazov N.R., Alimova I.S., Malykh V.A. Sentiment classification of reviews and twitter posts based on dictionaries. Russian Digital Libraries Journal, 2015, vol. 18, no. 3-4, pp. 138–162. (In Russian).
15. Panicheva P.V. ATEX: a rule-based sentiment analysis system processing texts in various topics. Computational Linguistics and Intellectual Technologies, 2013, iss. 12, vol. 2, pp. 101-113 (In Russian).
16. Kotelnikov E.V., Klekovkina M.V. Automatic text tonality analysis based on machine learning methods. Computational Linguistics and Intellectual Technologies, 2012. – iss. 11, vol. 2. – pp. 27-36. (In Russian).
17. Maltseva A.V., Makhnytkina O.V., Shilkina N.E., Lizunova I.A. Social media sentiment analysis with context space model. Communications in Computer and Information Science, 2020. – vol. 1135, pp. 399-412. doi:10.1007/978-3-030- 39296-3_29.
18. Aken B., Risch J., Krestel R., Loser A. Challenges for toxic comment classification: An in-depth error analysis. EMNLP, 2018. – pp. 33-42.
19. Voronina I.E., Goncharov V.A. Analysis of the emotional color of messages in social networks (for example, the “Vkontakte network”). Bulletin of the Voronezh State University. Series: System Analysis and Information Technologies. – 2015. – no. 4. – pp. 151-158. (In Russian).
20. Konig A.C., Brill E. Reducing the human overhead in text categorization. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2006. – pp. 598-603.
21. Lakshmi B.S., Raj P.S., Vikram R.R. Sentiment analysis using deep learning technique CNN with KMeans. International Journal of Pure and Applied Mathematics. – 2017. – vol. 114, no. 11. – pp. 47-57.
22. Kulagin D.I. Otkrytyj tonal'nyj slovar' russkogo yazyka KartaSlovSent // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog. – 2021. – №. 20. – S. 1106-1119.
23. Loukachevitch N., Levchik A. Creating a general Russian sentiment lexicon. Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’16). – 2016. – pp. 1171-1176.
24. Loukachevitch N., Dobrov B. RuThes linguistic ontology vs. Russian wordnets. Proceedings of the 7th Global Wordnet Conference. – 2014. – pp. 154-162.
25. Alexeeva S., Kolcov S., Koltsova O. Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media. Trudy XVIII ob”edinennoj konferencii «Internet i sovremennoe

obshchestvo» (IMS-2015) [Proceedings of the XVIII Joint Conference “Internet and Modern Society” (IMS2015)]. – 2015. – pp. 25-34. (In Russian).
26. Sokolova M., Bobicev V. Classification of emotion words in Russian and Romanian languages. Proceedings of the International Conference RANLP-2009. – 2009. – pp. 416-420.
27. Belyakov M.V. The analysis of news messages on the RF ministry of foreign affairs affairs website by the sentimental analysis (article 2). Bulletin of the Peoples’ Friendship University of Russia. Series: Theory of Language. Semiotics. Semantics. – 2016. – no. 4. – pp. 115-124. (In Russian).
28. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. Computational Linguistics and Intellectual Technologies. – 2015. – iss. 14, vol. 2. – pp. 3-13.
29. Lukashevich N.V., Rubtsova Y. V. SentiRuEval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. Computational Linguistics and Intellectual Technologies. – 2016. – iss. 15. – pp. 416-426.
30. Rubcova U.V. Building a text corpus for setting up a tone classifier. Software & Systems. – 2015. – no. 1(109), pp. 72–78. (In Russian).
31. Chetviorkin I., Braslavskiy P., Loukachevich N. Sentiment analysis track at ROMIP 2011. Computational Linguistics and Intellectual Technologies. – 2012. – iss. 11, vol. 2. – pp. 1-14.
32. Glazkova A.V. The evaluation of the proximity of text categories for solving electronic documents classification tasks. Bulletin of Tomsk State University. Management, Computer Engineering and Informatics. – 2015, no. 2(31). – pp. 18-25. doi:10.17223/19988605/31/2. (In Russian).
33. Perepelkina O., Kazimirova E., Konstantinova M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. Proceedings of 20th International Conference on Speech and Computer SPECOM-2018, Springer, Cham. – 2018. – pp. 501-510.
34. Zafar L., Afzal M.T., Ahmed U. Exploiting polarity features for developing sentiment analysis tool. CEUR-WS, 2017, vol. 1874, no. 4. Available at: http://ceur-ws.org/Vol1874/paper_4.pdf (accessed 2 May 2020).
35. Zvereva P. Sentiment-analysis of text (texts about Russia and the Russians from The New York Times). Bulletin of the Moscow State Regional University. Series: Linguistics. – 2014. – no. 5. – pp. 32-37. (In Russian).
36. Krivonogova S.A. Psychoemotional color of the text: theory and research methods. Materialy 68-j nauchnoj konferencii «Nauka YUURGU» [Materials of the 68th Scientific Conference “Science of the South Ural State University”]. – 2016. – vol. 100. – pp. 368–375. (In Russian).
37. Thelwall M. The heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. Cyberemotions, Springer, Cham. – 2017. – pp. 119-134.
38. Mayorov V., Andrianov I. MayAnd at SemEval-2016 Task 5: Syntactic and word2vec-based approach to aspect-based polarity detection in Russian. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). – 2016. – pp. 325-329.
39. Hercig T., Brychcin T., Svoboda L., Konkol M., Steinberger J. Unsupervised methods to improve aspect-based sentiment analysis in Czech. Computacion y Sistemas. – 2016. – vol. 20 (3), pp. 365-375. doi:10.13053/cys-20-3-2469
40. Hercig T., Brychcin T., Svoboda L., Konkol M. Uwb at semeval-2016 task 5: Aspect based sentiment analysis. SemEval-2016. – 2016. – pp. 342-349.

К.Е.Нурсакитов*, А.Т.Бекишев, С.К. Кумаргажанова, А.М.Уркумбаева
Д.Серікбаев атындағы Шығыс Қазақстан техникалық университеті,
070004, Өскемен, көш. Протозанова А.Қ., 69
e-mail*: nursakitov@bk.ru

ТАБИҒИ ТІЛДЕРДЕГІ МӘТІНДЕРДІҢ ТОНАЦИЯСЫН АНЫҚТАУ ӘДІСТЕРІН ҚАРАУ

Пайдаланушы пікірлеріндегі сезімді талдау тауарлар мен қызметтердің сапасын бағалау, хабарламалардағы эмоцияларды талдау және фишингтік жарнамаларды анықтау сияқты көптеген салаларда қолданылады. Орыс тіліндегі мәтіндік мәліметтерді сезімдік талдаудың көптеген әдістері бар, бірақ орыс тіліндегі мәтіндердің көңіл-күйін автоматты түрде талдау әлемнің басқа негізгі тілдеріне қарағанда әлдеқайда аз дамыған. Бұл мақала Қазақстанның киберкеңістігіндегі қауіпті контентті анықтаудың ақпараттық жүйесін

құру бойынша кеңірек зерттеудің бөлігі болып табылады. Бұл мақаланың мақсаты – орыс мәтіндеріндегі сезімді талдаудың әртүрлі тәсілдеріне аналитикалық шолу жасау және мәтінді жіктеу мәселесін шешудің заманауи әдістерін салыстыру. Сонымен қатар, мақала осы саладағы даму тенденцияларын анықтауға және одан әрі зерттеулерде пайдалану үшін ең жақсы алгоритмдерді таңдауға бағытталған. Шолу мәтіндік деректерді алдын ала өңдеудің, векторлаудың және мәтіндердің көңіл-күйін талдауға арналған машиналық классификацияның әртүрлі әдістерін қамтиды және осы тақырып бойынша бар дерекқорларды талдаумен аяқталады. Мақалада орыс тіліндегі мәтіндердегі көңіл-күйді талдаудың кейбір негізгі шешілмеген мәселелері айқындалып, алдағы уақытта жоспарланған зерттеулер талқыланады.

Түйін сөздер: *машиналық оқыту, қайталанатын нейрондық желілер, табиғи тілді өңдеу, мәтіндік сезім.*

К.Е.Нурсакитов*, А.Т. Бекишев, С.К. Кумаргажанова, А.М.Уркумбаева
Восточно-Казахстанский технический университет имени Д.Серикбаева,
070004, г. Усть-Каменогорск, ул. Протозанова А.К., 69
e-mail*: nursakitov@bk.ru

ОБЗОР МЕТОДОВ ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВ НА ЕСТЕСТВЕННЫХ ЯЗЫКАХ

Анализ настроений в комментариях пользователей находит применение во многих областях, таких как оценка качества товаров и услуг, анализ эмоций в сообщениях, обнаружение фишинговой рекламы. Существует множество методов анализа тональности текстовых данных на русском языке, но автоматический анализ тональности русскоязычных текстов разработан гораздо меньше, чем для других основных языков мира. Данная статья является частью более широкого исследования по созданию информационной системы обнаружения опасного контента в киберпространстве Казахстана. Цель данной статьи – дать аналитический обзор различных подходов к анализу тональности русскоязычных текстов и сравнить современные методы решения задачи классификации текстов. Кроме того, в статье ставится задача выявить тенденции развития в этой области и выбрать оптимальные алгоритмы для использования в дальнейших исследованиях. Обзор охватывает различные методы предварительной обработки текстовых данных, векторизации и машинной классификации для анализа тональности текстов и завершается анализом существующих баз данных по этой теме. В статье обозначены некоторые из основных нерешенных проблем при анализе тональности русскоязычных текстов и обсуждаются планируемые дальнейшие исследования.

Ключевые слова: *машинное обучение, рекуррентные нейронные сети, обработка естественного языка, тональность текста.*

Information about the authors

Kuanysh Nursakitov* – doctoral student of the "School of Information Technologies and Intelligent Systems"; East Kazakhstan Technical University. D. Serikbaev of the city of Ust-Kamenogorsk, Republic of Kazakhstan; e-mail: nursakitov@bk.ru.

Askar Bekishev – doctoral student of the "School of Information Technologies and Intelligent Systems"; East Kazakhstan Technical University. D. Serikbaev of the city of Ust-Kamenogorsk, Republic of Kazakhstan; e-mail: bekishev@bk.ru.

Saule Kumargazhanova – Candidate of Technical Sciences: Dean of the "School of Information Technologies and Intelligent Systems"; East Kazakhstan Technical University. D. Serikbaev of the city of Ust-Kamenogorsk, Republic of Kazakhstan; e-mail: skumargazhanova@gmail.com.

Aliya Urkumbayeva – Candidate of Technical Sciences, Senior Lecturer of the "School of Information Technologies and Intelligent Systems"; East Kazakhstan Technical University. D. Serikbaev of the city of Ust-Kamenogorsk, Republic of Kazakhstan; e-mail: urkumbaeva@mail.ru

Авторлар туралы мәліметтер

Қуаныш Ерлибекович Нұрсакитов* – «Ақпараттық технологиялар және зияткерлік жүйелер мектебінің» докторанты; Шығыс Қазақстан техникалық университеті. Қазақстан Республикасы Өскемен қаласының Д. Серікбаев; e-mail: nursakitov@bk.ru.

Аскар Тлеуханович Бекішев – «Ақпараттық технологиялар және зияткерлік жүйелер мектебінің» докторанты; Шығыс Қазақстан техникалық университеті. Қазақстан Республикасы Өскемен қаласының Д. Серікбаев; e-mail: bekishev@bk.ru.

Сауле Кумаргажановна Құмарғажанова – техника ғылымдарының кандидаты, «Ақпараттық технологиялар және интеллектуалды жүйелер мектебі» деканы; Шығыс Қазақстан техникалық университеті. Қазақстан Республикасы Өскемен қаласының Д. Серікбаев; e-mail: skumargazhanova@gmail.com.

Алия Муратовна Уркумбаева – техника ғылымдарының кандидаты, «Ақпараттық технологиялар және интеллектуалды жүйелер мектебі» кафедрасының аға оқытушысы; Шығыс Қазақстан техникалық университеті. Қазақстан Республикасы Өскемен қаласының Д. Серікбаев; e-mail: urkumbaeva@mail.ru.

Сведения об авторах

Куаныш Ерлибекович Нурсакитов* – докторант «школы информационных технологий и интеллектуальных систем»; Восточно-Казахстанский технический университет им. Д. Серикбаева города Усть-Каменогорск, Республика Казахстан; e-mail: nursakitov@bk.ru.

Аскар Тлеуханович Бекишев – докторант «школы информационных технологий и интеллектуальных систем»; Восточно-Казахстанский технический университет им. Д. Серикбаева города Усть-Каменогорск, Республика Казахстан; e-mail: bekishev@bk.ru.

Сауле Кумаргажановна Кумаргажанова – кандидат технических наук: декан «школы информационных технологий и интеллектуальных систем»; Восточно-Казахстанский технический университет им. Д. Серикбаева города Усть-Каменогорск, Республика Казахстан; e-mail: skumargazhanova@gmail.com.

Алия Муратовна Уркумбаева – кандидат технических наук, старший преподаватель «школы информационных технологий и интеллектуальных систем»; Восточно-Казахстанский технический университет им. Д. Серикбаева города Усть-Каменогорск, Республика Казахстан; e-mail: urkumbaeva@mail.ru.

Material received on 13.03.2023.