UDC 004.051 **DOI: 10.20998/2079-0023.2019.02.13**

N. V. BORYSOVA, K. V. MELNYK

EFFICIENCY ESTIMATION OF METHODS FOR SENTIMENT ANALYSIS OF SOCIAL NETWORK MESSAGES

The results of effectiveness evaluating of machine learning methods for sentiment analysis of social network messages are presented in this paper. The importance of the sentiment analysis problem as one of the important tasks of natural language processing in general and textual information processing in particular is substantiated. A review of existing methods and software for sentiment analysis are made. The choice of classifiers for sentiment analysis of texts for this research is substantiated. The principles of functioning of a Naïve Bayesian Classifier and classifier based on a recurrent neural network are described. Classifiers were sequentially trained in two corpuses: first, in the RuTweetCorp corpus, the corpus of short messages from the social network Twitter, and then on the Slang corpus, the corpus of messages from social networks Facebook and Instagram and posts from the Pikabu website, second corpus have been marked up the tonality of slang words. Information about the tonality of slang words was taken from the youth slang dictionary obtained as a result of the survey of users. The separation of texts by tonality was carried out into three classes: positive, negative and neutral. The efficiency of these classifiers was evaluated. Efficiency evaluation was carried out according to standard metrics Recall, Precision, F-measure, Accuracy. For the naive Bayesian classifier, after training on the first corpus, the following metric values were obtained: Recall = 0,853; Precision = 0,869; F-measure = 0,861; Accuracy = 0,855; and after training on the second corpus such values were obtained: Recall = 0,948; Precision = 0,975; F-measure = 0,961; Accuracy = 0,960. For the classifier based on a recurrent neural network, after training on the first corpus, the following metric values were obtained: Recall = 0,870; Precision = 0,878; F-measure = 0,874; Accuracy = 0,861; and after training on the second corpus such values were obtained: Recall = 0,965; Precision = 0,982; F-measure = 0,973; Accuracy = 0,973. These results prove that additional training on the second corpus increased the efficiency of classifiers by 10–11%.

Keywords: sentiment analysis, social networks messages analysis, machine learning, text classification, naïve Bayesian classification, recurrent neural network, efficiency estimation

Н. В. БОРИСОВА, К. В. МЕЛЬНИК

ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДІВ СЕНТИМЕНТ-АНАЛІЗУ ПОВІДОМЛЕНЬ СОЦІАЛЬНИХ МЕРЕЖ

У роботі представлено результати оцінки ефективності методів машинного навчання для сентимент-аналізу повідомлень соціальних мереж. Обґрунтовано актуальність задачі сентимент-аналізу як однієї з важливих задач обробки природної мови взагалі та обробки текстової інформації зокрема. Проведено огляд існуючих методів сентимент-аналізу та програмних продуктів, що вирішують цю задачу. Обґрунтовано вибір класифікаторів для сентимент-аналізу текстів у межах дослідження. Описано принципи роботи наївного байєсівського класифікатора та класифікатора на основі рекурентної нейронної мережі. Класифікатори було послідовно навчено на двох корпусах: спочатку на корпусі RuTweetCorp – корпусі коротких повідомлень соціальної мережі Twitter, а потім на корпусі Slang corpus – корпусі повідомлень соціальних мереж Facebook та Instagram і постів з сайту Pikabu, у якому розмічено тональність сленгових слів. Інформацію про тональність сленгових слів було взято із словника молодіжного сленгу, отриманого у результаті опитування користувачів. Розподіл текстів за тональністю здійснювався на три класи: позитивні, негативні й нейтральні. Проведено оцінку ефективності роботи цих класифікаторів. Оцінка ефективності здійснювалась за стандартними метриками Recall, Precision, F-measure, Accuracy. Для наївного байєсівського класифікатора після навчання на першому корпусі були отримані наступні значення метрик: Recall = 0,853; Precision = 0,869; F-measure = 0,861; Accuracy = 0,855; а після навчання на другому корпусі такі значення: Recall = 0,948; Precision = 0,975; F-measure = 0,961; Accuracy = 0,960. Для класифікатора на основі рекурентної нейронної мережі після навчання на першому корпусі були отримані наступні значення метрик: Recall = 0,870; Precision = 0,878; F-measure = 0,874; Accuracy = 0,861; а після навчання на другому корпусі такі значення: Recall = 0,965; Precision = 0,982; F-measure = 0,973; Accuracy = 0,973 Отримані результати довели, що додаткове навчання на другому корпусі підвищило ефективність роботи класифікаторів на 10–11%.

Ключові слова: сентимент-аналіз, аналіз повідомлень соціальних мереж, машинне навчання, класифікація текстів, наївний байєсівський класифікатор, рекурентна нейронна мережа, оцінка ефективності

Н. В. БОРИСОВА, К. В. МЕЛЬНИК

ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДОВ СЕНТИМЕНТ-АНАЛИЗА СООБЩЕНИЙ СОЦИАЛЬНЫХ СЕТЕЙ

В работе представлены результаты оценки эффективности методов машинного обучения для сентимент-анализа сообщений социальных сетей. Обоснована актуальность задачи сентимент-анализа как одной из важных задач обработки естественного языка вообще и обработки текстовой информации в частности. Проведен обзор существующих методов сентимент-анализа и программных продуктов, решающих эту задачу. Обоснован выбор классификаторов для сентимент-анализа текстов в рамках исследования. Описаны принципы работы наивного байесовского классификатора и классификатора на основе рекуррентной нейронной сети. Классификаторы были последовательно обучены на двух корпусах: сначала на корпусе RuTweetCorp – корпусе коротких сообщений социальной сети Twitter, а затем на корпусе Slang corpus – корпусе сообщений социальных сетей Facebook и Instagram и постов с сайта Pikabu, в котором размечена тональность сленговых слов. Информация о тональности сленговых слов была взята из словаря молодежного сленга, полученного в результате опроса пользователей. Разделение текстов по тональности осуществлялось на три класса: позитивные, негативные и нейтральные. Проведена оценка эффективности работы этих классификаторов. Оценка эффективности осуществлялась по стандартным метрикам Recall, Precision, F-measure, Accuracy. Для наивного байесовского классификатора после обучения на первом корпусе были получены следующие значения метрик: Recall = 0,853; Precision = 0,869; F-measure = 0,861; Accuracy = 0,855; а после обучения на втором корпусе такие значения: Recall = 0,948; Precision = 0,975; F-measure = 0,961; Accuracy = 0,960. Для классификатора на основе рекуррентной нейронной сети после обучения на первом корпусе были получены следующие значения метрик: Recall = 0,870; Precision = 0,878; F-measure = 0,874; Accuracy = 0,861; а после обучения на втором корпусе такие значения: Recall = 0,965; Precision = 0,982; F-measure = 0,973; Accuracy = 0,973 Полученные результаты доказывают, что дополнительное обучение на втором корпусе повысило эффективность работы классификаторов на 10–11%.

Ключевые слова: сентимент-анализ, анализ сообщений социальных сетей, машинное обучение, классификация текстов, наивный байесовский классификатор, рекуррентная нейронная сеть, оценка эффективности

Introduction. The task of analyzing the tonality of the text or sentiment analysis is the task of determining the emotional attitude of the author to a certain object, which described in the text. This task is one of the most relevant NLP tasks. The sentiment analysis is used for assessing the quality of goods and services according to the Internet user reviews, for identifying the criminally significant content, for determent the authorship of texts, for predicting various economic indicators, for generating of texts with a pre-established emotional coloring. The amount of information in electronic form increases exponentially. So, it is not possible to analyze it manually, therefore, there is a need for automatic methods and tools of analyzing textual information, including methods and tools for automated sentiment analysis.

Last researches and publications analysis. The analytical review of different sources has showed great interest of researchers to the task of sentiment analysis [1, 3, 7–9, 11]. In a basic this task is the task of texts classifying. The result of the task is a set of texts, where texts or elements are divided into two (positive, negative), three (positive, neutral, negative), five (positive, rather positive, neutral, rather negative, negative) or more classes. There are many methods, which can be used for resolving this task. It can be divided into several groups. The first group includes methods based on rules and dictionaries that use pre-compiled emotive dictionaries and linguistic rules for searching of emotive words. The first step of the process of assigning the text to definite class is a search of words from emotive dictionaries. The second step is assigning the all found word its tonality or weight from the dictionary. Then the overall tonality of the text is calculated by summing the tonality values of each found word. The second group includes machine learning methods with a teacher, which used a pre-trained classifier to determine the tonality of new texts. The classifier is trained on a specially selected collection of texts with definite type of tonality. The third group includes machine learning methods without a teacher. In this case, the methods determine the tonality of the terms that have the greatest weight. The frequency of these terms should be greatest in certain text and at the same time it should be present in a small number in the texts throughout the collection. Then the tonality of the entire text is determined by using the tonality of the terms. The combination of different methods from different groups is perspective way to obtain a better result.

The aforementioned methods are widely used in appropriate software for text sentiment analysis, such as «Analytical Courier» [13], «RCO Fact Extractor SDK» [10], «VAAL» [14], «Eureka Engine» [3], SentiStrength [12], etc. They have quite good functionality, but are not without some drawbacks, especially related to the analysis of inflected languages with a rich morphology.

Therefore, **the purpose of the work** is to verify the efficiency of various methods of social network messages sentiment analysis.

The main material. The task of sentiment analysis of social networks messages is basically the same as a classification task. Let's consider this task in the context of the separation of texts into three categories: positive tonality, neutral and negative. Formally, this task can be represented as follows: if we denote by $W = \{w_1, \dots, w_n\}$ a set of emotionally colored words and phrases, and $S = \{s_1, s_2, s_3\}$ is a set of three classes of tonality of texts, then the task of determining the emotional attitude of the author to a certain object, event or the process of the real world looks like this: $f: W \to S$ is to find a mapping of one set to another.

In this research, to solve the problem of sentiment analysis of texts in the proposed formulation, two approaches were analyzed and their efficiency was estimated, namely, a Naïve Bayesian Classifier and a classifier based on recurrent neural network. The Naïve Bayesian Classifier was chosen because it trains and works faster than all other classifiers, and at the same time it solves the problem quite effectively. A recurrent neural network, in comparison with other types of neural networks, is best suited for working with texts, since it can use its internal memory to process sequences of arbitrary length, it can process output data of arbitrary length, new information in it can be used to obtain the following state of hidden layers, it contains feedbacks that allow to save information.

A detailed algorithm for solving the classification problem by using the Naïve Bayesian Classifier is considered in [5]. Let's consider using of the Naïve Bayesian Classifier for sentiment analysis. Let's introduce the necessary notation. If T is a social media training message text template, then T_j is a j-th text from training template T. Previously it was indicated that w_i is a presence of definite word or word combination in set W. Denote the presence or absence of w_i in the T_j as w_{ij}

$$w_{ij} = \begin{cases} 1, & w_i \in T_j \\ 0, & w_i \notin T_j \end{cases}.$$

Then x_{iz} is a number of appearance of w_i in z-th text tonality class, where $z = \overline{1,3}$ is a number of text tonality class $s_z \in S$. Let's s_{zj} is output class of T_j text. Denote number of appearance of z-th text tonality class in training template T as y_z.

Taking into account the introduced notation, the classification algorithm for sentiment analysis using the Bayesian Classifier has the following steps.

I. Training of the Naïve Bayesian Classifier:

1. Calculate the number of appearance of w_i for each text tonality class separately

$$x_{iz} = \sum_{i \in W,\ j \in T} w_{ij}\,,\ z = \overline{1,3}.$$

2. Calculate number of appearance of s_z in training template T

$$y_z = \sum_{j \in T} s_{zj}\,,\ z = \overline{1,3}.$$

3. Calculate conditional probability $P(^{w_i}/_{s_z})$ of occurrence w_i in z-th text tonality class

$$P(^{w_i}/_{s_z}) = \frac{x_{iz}}{\sum_{z=1}^{3} x_{iz}}.$$

4. Calculate probability $P(s_z)$ of T_j-th text assignment to the z-th text tonality class

$$P(s_z) = \frac{y_z}{\sum_{z=1}^{3} y_z}.$$

II. Using of the Naïve Bayesian Classifier:

1. Calculate conditional probabilities $P\left(^{s_z}/_{\{w_{ij+1}\}}\right)$ of T_{j+1}-th text

$$P\left(^{s_z}/_{\{w_{ij+1}\}}\right) = P(s_z) * \prod_z P(^{w_i}/_{s_z}).$$

2. Define output class s_{zj+1} of the T_{j+1}-th text. Denote R_z as the conditional probability of the output class of T_{j+1}-th text

$$R_z = P\left(^{s_z}/_{\{w_{ij+1}\}}\right),$$

$$s_{zj+1} = \arg\max_{z=\overline{1,3}} R_z .$$

As mentioned earlier, in addition to the Naïve Bayesian Classifier, the efficiency of the classifier based on recurrent neural network was also evaluated in this paper. We used an architecture for neural network called a simple recurrent neural network or Elman network [4]. This is the recurrent neural network version that very easy to implement and train. The network has an input layer x, hidden layer s (also called context layer or state) and output layer y. Input to the network in time t is $x(t)$, output is denoted as $y(t)$, and $s(t)$ is state of the network (hidden layer). Input vector $x(t)$ is formed by concatenating vector $w$representing current word, and output from neurons in context layer s at time $t-1$. Then input, hidden and output layers are computed as follows:

$$x(t) = w(t) + s(t-1),$$

$$s_j(t) = f\left(\sum_i x_i(t) u_{ij}\right),$$

$$y_k(t) = g\left(\sum_j s_j(t) v_{kj}\right),$$

where $f(z)$ is sigmoid activation function:

$$f(z) = \frac{1}{1+e^{-z}}$$

and $g(z)$ is softmax function

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

For initialization, $s(0)$ can be set to vector of small values. In the next time steps, $s(t+1)$ is a copy of $s(t)$. Input vector $x(t)$ represents word in time t encoded using $1-\text{of}-N$ coding and previous context layer – size of vector x is equal to size of vocabulary V plus size of context layer.

Networks are trained in several epochs, in which all data from training corpus are sequentially presented. Weights are initialized to small values. After each epoch, the network is tested on validation data. If log-likelihood of validation data increases, training continues in new epoch. If no significant improvement is observed, learning rate is halved at start of each new epoch. After there is again no significant improvement, training is finished.

Output layer $y(t)$ represents probability distribution of next word given previous word $w(t)$ and context $s(t-1)$. Softmax ensures that this probability distribution is valid: $y_m(t) => 0$ for any word m and $\sum_k y_k(t) = 1$.

At each training step, error vector is computed according to cross entropy criterion and weights are updated with the standard backpropagation algorithm:

$$error(t) = desired(t) - y(t),$$

where $desired(t)$ is a vector using $1-\text{of}-N$ coding representing the word that should have been predicted in a particular context and $y(t)$ is the actual output from the network [6].

The Naïve Bayesian Classifier and the classifier based on the recurrent neural network were trained on the same data set – the Russian-language corpus of short texts RuTweetCorp [18], consisting of 114 911 positive, 111 923 negative and 107 990 neutral entries for time period from the end of November 2013 to the end of February 2014. Each text in the corpus has the following attributes: publication date; author's name; Tweet text; the class to which the text belongs (positive, negative, neutral); the number of messages added to favorites; the number of retweets (the number of copies of this message by other users); number of friends of the user; the number of users who have this user in friends (number of followers); the number of lists the user is in [18, 20]. After training, the sentiment analysis of new texts was made by both classifiers. The results of classifiers efficiency evaluation after training on the RuTweetCorp corpus are presented in the Table 1.

Also in the research, it was decided to test the hypothesis that the efficiency of classifiers will increase if they are additionally trained on the social networks messages corpus [17], in which the tonality of slang words have been marked up. A similar hypothesis but with another formulation was provided in the [15]. Taking into account the tonality of slang words is important, since at the present stage of the development of the linguistic culture of society, the use of slang words is more and more noticeable, they enter both the everyday speech of almost all segments of the population and the media space,

especially the Internet media space. In addition, according to numerous linguistic studies, slang words and expressions are used to create the effect of novelty, unusualness; transmission of a certain mood of the speaker; giving the statement concreteness, liveliness, expressiveness, brevity, imagery, i.e. it can be fully used for sentiment analysis of texts.

Let's called the second corpus as Slang corpus. It consists of social networks Facebook and Instagram messages as well as the messages and posts from Pikabu web-site. It contains approximately 150000 words [17]. The emotional tone was tagged for each slang word in Slang corpus. Information about slang words' emotional tones were taken from youth slang dictionary [16], that contains approximately five thousand slang words (1493 positive, 1344 negative and 2141 neutral words). The results of classifiers efficiency evaluation after additional training on the Slang corpus are also presented in the table 1.

The experiment steps are represented in Figure 1 in IDEF0 notation. The functional modeling of the training process by using IDEF0 notation consists of two stages. The first stage shows the process of training the classifier on the RuTweetCorp corpus. After that, the calculation of efficiency is carried out. At the second step, the slang corpus is using for classifier training. Then numerical calculations and analysis of the results are carried out.

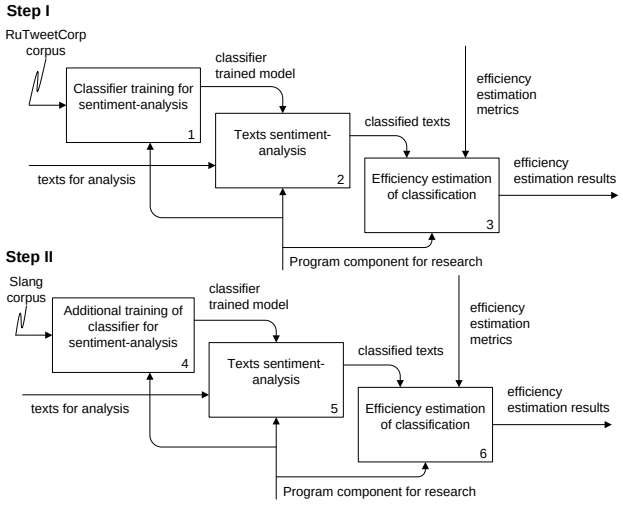

Fig. 1. The experiment steps in IDEF0 notation

As you can see, the input data is marked up corpuses, and the result is the numerical values of the Recall, Precision, Accuracy, F-measure metrics to estimate the classification efficiency.

Results and discussion. To assess the quality of the obtained classification results, generally accepted metrics were used: $Recall$, $Precision$, $F-measure$, $Accuracy$. For the calculation of metrics the values of the following parameters were calculated:

- TP is the number of true positive results;
- TN is the number of true negative results;
- FP is the number of false positive results;
- FN is the number of false negative results.

$Precision$ is the proportion of objects classified as X that really belong to class X:

$$Precision = \frac{TP}{TP + FP}.$$

$Recall$ is the proportion of all objects of class X classified as belonging to class X:

$$Recall = \frac{TP}{TP + FN}.$$

$F-measure$ is the harmonic mean between $Precision$ and $Recall$:

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

$Accuracy$ is the proportion of right classified objects in the all classified objects:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

The results of classifiers efficiency evaluation after training on two corpuses are presented in the Table 1. In this table NBC means Naïve Bayesian Classifier, and RNNC means Classifier based on Recurrent Neural Network.

Table 1 – Efficiency estimation of classification results

Metrics	Corpus RuTweetCorp		Slang corpus	
	NBC	RNNC	NBC	RNNC
Recall	0,853	0,870	0,948	0,965
Precision	0,869	0,878	0,975	0,982
F-measure	0,861	0,874	0,961	0,973
Accuracy	0,855	0,861	0,960	0,973

As the efficiency estimation results analysis shows with additional training on the second Slang corpus the efficiency of classifiers increased by 10–11%, which confirms the research hypothesis proposed earlier.

Conclusions. A comparison of the efficiency estimation results of the Naïve Bayesian Classifier with the results obtained by other researchers on the RuTweetCorp corpus [20] showed that the discrepancies are insignificant. However, it is not possible to compare the efficiency of the classifier based on recurrent neural network with similar ones due to the lack of references to such researches with the RuTweetCorp corpus.

References

1. Ameur H., Jamoussi S., Hamadou A.B. A New Method for Sentiment Analysis Using Contextual Auto-Encoders. *Journal of Computer Science and Technology*. 2018. Volume 33, issue 6. P. 1307–1319. DOI: https://doi.org/10.1007/s11390-018-1889-1.
2. *Eureka Engine*. URL: http://eurckacngine.ru/ru/description (access date: 15.09.2019).
3. Huang M., Zhuang F., Zhang X. et al. Supervised representation learning for multi-label classification. *Machine Learning*, 2019. Volume 108, issue 5. P. 747–763. DOI: https://doi.org/10.1007/s10994-019-05783-5.
4. Elman J. L. Finding Structure in Time. *Cognitive Science*. 1990. Volume 14, issue 2. P. 179–211.

5. Melnyk K. V., Borysova N. V. Improving the quality of credit activity by using scoring model. *Radio Electronics, Computer Science, Control*. 2019. Volume 2. P. 60–70. DOI 10.15588/1607-3274-2019-2-7 . e-ISSN 1607-3274.
6. Mikolov T., Karafiat M., Burget L., Cernocky J., Khudanpur S. Recurrent neural network based language model. *Proceedings 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Chiba, Japan, 2010. P. 1045–1048.
7. Nguyen-Trang T., Vo-Van T. A new approach for determining the prior probabilities in the classification problem by Bayesian method. *Advances in Data Analysis and Classification*. 2017. Volume 11, issue 3. P. 629–643. DOI: https://doi.org/10.1007/s11634-016-0253-y.
8. Pang B., Lee L., Vaithyanathan Sh. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02), Association for Computational Linguistics*. *Volume 10*. 2002. P. 79–86. DOI: https://doi.org/10.3115/1118693. 1118704.
9. Rahimi Z., Noferesti S., Shamsfard M. Applying data mining and machine learning techniques for sentiment shifter identification. *Language Resources and Evaluation*. 2019. Volume 53, issue 2. P. 279–302. DOI: https://doi.org/10.1007/s10579-018-9432-0 .
10. *RCO Fact Extractor SDK*. URL: http://www.rco.ru/?page_id=3554. (access date: 15.09.2019).
11. Rubtsova Y. Automatic Term Extraction for Sentiment Classification of Dynamically Updated Text Collections into Three Classes. *Proceedings of International Conference on Knowledge Engineering and the Semantic Web (KESW 2014), Communications in Computer and Information Science*. *Volume 468*. P. 140–149. DOI: https://doi.org/10.1007/978-3-319-11716-4_12.
12. *SentiStrength – sentiment strength detection in short texts*. URL: http://sentistrength.wlv.ac.uk/#About (access date: 15.09.2019).
13. *System «Analytical Courier»*. URL: http://www.iteco.ru/solutions/business_intclligence_products/analytical_courier (access date: 15.09.2019).
14. *VAAL project*. URL: http://www.vaal.ru (access date: 15.09.2019).
15. Wu L., Morstatter F., Liu H. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*. 2018. Volume 52, issue 3. P. 839–852. DOI: https://doi.org/10.1007/s10579-018-9416-0
16. Борисова Н. В., Ніфтілін В. В. Автоматизоване створення електронного словника. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXV Міжнародної науково-практичної конференції MicroCAD-2017*. Ч. І. Харків: НТУ «ХПІ», 2017. С. 32.
17. Борисова Н. В., Ніфтілін В. В. Застосування методів корпусної лінгвістики для дослідження особливостей використання сучасного молодіжного сленгу. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXVI міжнародної науково-практичної конференції MicroCAD-2018*. Ч 1. Харків: НТУ «ХПІ», 2018. С. 27.
18. *Корпус коротких текстов RuTweetCorp*. URL: http://study.mokoron.com (access date: 15.09.2019).
19. Романов А. В., Васильева М. И., Куртукова А. В., Мещеряков Р. В. Анализ тональности текстов с использованием методов машинного обучения. *Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. CEUR Workshop Proceedings. Volume 2233*. Saint Petersburg, Russia, 2017. P. 86–95.
20. Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора. *Программные продукты и системы*, 2015. № 1 (109). С. 72–78. DOI: 10.15827/0236-235X.109.072-078.

References (transliterated)

1. Ameur H., Jamoussi S., Hamadou A.B. A New Method for Sentiment Analysis Using Contextual Auto-Encoders. *Journal of Computer Science and Technology*. 2018, vol. 33, issue 6, pp. 1307–1319. DOI: https://doi.org/10.1007/s11390-018-1889-1.
2. *Eureka Engine*. Available at: http://eurckacngine.ru/ru/description (accessed 15.09.2019).
3. Huang M., Zhuang F., Zhang X. et al. Supervised representation learning for multi-label classification. *Machine Learning*. 2019, vol. 108, issue 5, pp. 747–763. DOI: https://doi.org/10.1007/ s10994-019-05783-5.
4. Jeffrey L. Elman. Finding Structure in Time. Cognitive Science. 1990, vol. 14, issue 2, pp. 179–211.
5. Melnyk K. V., Borysova N. V. Improving the quality of credit activity by using scoring model. *Radio Electronics, Computer Science, Control*. 2019, vol. 2, pp. 60–70. DOI 10.15588/1607-3274-2019-2-7 . e-ISSN 1607-3274.
6. Mikolov T., Karafiat M., Burget L., Cernocky J., Khudanpur S. Recurrent neural network based language model. *Proceedings 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*. Makuhari, Chiba, Japan, 2010, pp. 1045–1048.
7. Nguyen-Trang T., Vo-Van T. A new approach for determining the prior probabilities in the classification problem by Bayesian method. *Advances in Data Analysis and Classification*. 2017, volume 11, issue 3, pp. 629–643. DOI: https://doi.org/10.1007/s11634-016-0253-y.
8. Pang B., Lee L., Vaithyanathan Sh. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02), Association for Computational Linguistics*. *Vol. 10*. 2002, pp. 79–86. DOI: https://doi.org/10.3115/1118693. 1118704.
9. Rahimi Z., Noferesti S., Shamsfard M. Applying data mining and machine learning techniques for sentiment shifter identification. *Language Resources and Evaluation*, 2019, vol. 53, issue 2, pp. 279–302. DOI: https://doi.org/10.1007/s10579-018-9432-0 .
10. *RCO Fact Extractor SDK*. Available at: http://www.rco.ru/?page_id=3554. (accessed 15.09.2019).
11. Rubtsova Y. Automatic Term Extraction for Sentiment Classification of Dynamically Updated Text Collections into Three Classes. *Proceedings of International Conference on Knowledge Engineering and the Semantic Web (KESW 2014), Communications in Computer and Information Science*. *Vol. 468*. Pp. 140–149. DOI: https://doi.org/10.1007/978-3-319-11716-4_12.
12. *SentiStrength – sentiment strength detection in short texts*. Available at: http://sentistrength.wlv.ac.uk/#About (accessed 15.09.2019).
13. *System «Analytical Courier»*. Available at: http://www.iteco.ru/solutions/business_intclligence_products/analytical_courier (accessed 15.09.2019).
14. *VAAL project*. Available at: http://www.vaal.ru (accessed 15.09.2019).
15. Wu L., Morstatter F., Liu H. SlangSD: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*. 2018, vol. 52, issue 3, pp. 839–852. DOI: https://doi.org/10.1007/ s10579-018-9416-0.
16. Borysova N. V, Niftilin V. V. Avtomatyzovane stvorennia elektronnogo slovnyka [Automated creation of electronic dictionary]. *Informaciyni technologii: nauka, technika, technologiia, osvita, zdorov'ia: tezy dopovidei XXV Mizhnarodnoi naukovo-practychnoi konferencii MicroCAD-2017. Ch. I* [Proceedings of XXV International scientific-practical conference in Information technologies: science, engineering, technology, education, health MicroCAD-2017. Part I.]. Kharkiv: NTU "KhPI", 2017, p. 32.
17. Borysova N. V, Niftilin V. V. Zastosuvaniia metodiv korpusnoi lingvistiki dlia doslidzhennia osoblyvostei vykorystannia suchasnogo molodizhnogo slengu [Using of corpus linguistics methods to study the features of using modern youth slang]. *Informaciyni technologii: nauka, technika, technologiia, osvita, zdorov'ia: tezy dopovidei XXV Mizhnarodnoi naukovo-practychnoi konferencii MicroCAD-2018. Ch. I* [Proceedings of XXV International scientific-practical conference in Information technologies: science, engineering, technology, education, health MicroCAD-2018. Part I.]. Kharkiv: NTU "KhPI", 2018, p. 27.
18. *Korpus korotkih tekstov RuTweetCorp* [Short texts corpus RuTweetCorp]. Available at: http://study.mokoron.com (accessed 15.09.2019).
19. Romanov A. V., Vasilieva M. I., Kurtukova A. V., Meshcheriakov R. V. Analiz tonalnosti tekstov s ispolzovaniem metodov mashinnogo obucheniia [Sentiment Analysis of Text Using Machine Learning Techniques]. *Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. CEUR Workshop Proceedings. Vol.-2233*. Saint Petersburg, Russia, 2017, pp. 86–95.

20. Rubtsova Yu. V. Postroenie korpusa tekstov dlia nastroyki tonovogo klassifikatora [Constructing a corpus for sentiment classification training]. *Programnye produkty i sistemy* [Program products and systems]. 2015, no. 1 (109), pp. 72–78. DOI: 10.15827/0236-235X.109.072-078.

Received 25.09.2019

Відомості про авторів /Сведения об авторах/ About the Authors

Борисова Наталя Володимирівна (***Borysova Natalia Volodymyrivna***) – кандидат технічних наук, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри інтелектуальних комп'ютерних систем, м. Харків, Україна; ORCID: https://orcid.org/0000-0002-8834-2536; e-mail: borysova.n.v@gmail.com

Мельник Каріна Володимирівна (***Melnyk Karina Volodymyrivna***) – кандидат технічних наук, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри програмної інженерії та інформаційних технологій управління, м. Харків, Україна; ORCID: https://orcid.org/0000-0001-9642-5414; e-mail: karina.v.melnyk@gmail.com

Борисова Наталья Владимировна – кандидат технических наук, Национальный технический университет «Харьковский политехнический институт», доцент кафедры интеллектуальных компьютерных систем; г. Харьков, Украина; ORCID: https://orcid.org/0000-0002-8834-2536; e-mail: borysova.n.v@gmail.com

Мельник Карина Владимировна – кандидат технических наук, Национальный технический университет «Харьковский политехнический институт», доцент кафедры программной инженерии и информационных технологий управления; г. Харьков, Украина; ORCID: ORCID: https://orcid.org/0000-0001-9642-5414; e-mail: karina.v.melnyk@gmail.com

Borysova Natalia Volodymyrivna – Candidate of Engineering Sciences, National Technical University "Kharkiv Polytechnic Institute", Associate Professor, Department of Computer Science and Intellectual Property; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0002-8834-2536; e-mail: borysova.n.v@gmail.com

Melnyk Karina Volodymyrivna – Candidate of Engineering Sciences, National Technical University "Kharkiv Polytechnic Institute", Associate Professor, Department of Software Engineering and Management Information Technology; Kharkiv, Ukraine; ORCID: https://orcid.org/0000-0001-9642-5414; e-mail: karina.v.melnyk@gmail.com

УДК 004.272.26: 004.272.34: 519.876.5 **DOI: 10.20998/2079-0023.2019.02.14**

С. В. ШЕВЧЕНКО, В. О. ГУЖВА, В. Д. МАЛИШ, І. Ю. МОРКВА

ОБҐРУНТУВАННЯ ПОПЕРЕДНЬОГО ВИБОРУ АРХІТЕКТУРИ СИСТЕМИ ОБРОБКИ ДАНИХ З ВИКОРИСТАННЯМ НЕЧІТКОЇ ЛОГІКИ

Метою роботи є формування підходу до попереднього обґрунтування вибору типу архітектури системи обробки даних і управління. Архітектура системи являє собою способи побудови та організації її функціонування в процесі виконання програм обробки даних і управління. Якість архітектури може бути розглянуто з позицій прийнятих критеріїв ефективності таких як, наприклад, продуктивність, обсяги ресурсів, вартість обробки та інші. Вихідними даними для прийняття рішень по вибору кращою архітектури є характеристики даних задач, алгоритми обробки, характеристики прийнятних типів архітектури обчислювальних пристроїв, умови і вимоги до організації обчислювальних процесів і процесів управління, процедури обробки, їх характеристики і параметри, особливості програмного середовища, інструментальних засобів розробки і модифікації програмних рішень. Наявність невизначеності, викликаної майбутніми аспектами функціонування системи обробки даних і умовами її використання, а також зовнішніми і внутрішніми факторами, що постійно змінюються, призводить до необхідності використання підходів формування архітектури системи обробки даних з позицій зменшення ризику прийняття необґрунтованих рішень. Тому виникають потреби в обробці даних у складі робочого навантаження, яке змінюється у часі, що проявляється як у сукупності задач обробки та їх вихідних даних, так і в необхідних процедурах обробки. Ці умови формують середовище обробки даних, для якого може бути поставлена у відповідність система обробки з адекватною архітектурою. Ступінь адекватності архітектури такої системи може бути оцінена з позицій обраних критеріїв і рівнів їх узгодження. Варіанти архітектури системи, що відповідають узгодженим рішенням, складають підмножину, яка надає обґрунтовані варіанти вибору рішень, що можуть прийматися з оцінками ефективності. З огляду на зростаючий інтерес замовників до побудови обчислювальних систем на основі хмарних технологій, обґрунтування та вибір архітектури системи обробки даних з використанням послуг хмарних обчислень набуває особливої актуальності. Підготовка подібних систем до застосування може займати кілька хвилин. Тому для поліпшення якості обґрунтування попереднього вибору архітектури системи обробки даних пропонується використовувати процедури апарату нечіткої логіки. Для ілюстрації підходу пропонується приклад чисельних розрахунків та аналіз отриманих результатів.

Ключові слова: архітектура, комп'ютерна система, обробка даних, критерії, нечітка логіка, алгоритм.