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 choice 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 naïve 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%.

Keywords: sentiment analysis, social networks messages analysis, machine learning, text classification, naïve Bayesian classification, recurrent neural network, efficiency estimation
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 is 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 determining 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 shown 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 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 words 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 they 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 $W = \{w_1, ..., w_n\}$ 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: $f: W \rightarrow S$ 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 $z = \frac{1}{3}$ 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:

1. Training of the Naïve Bayesian Classifier:
   1. Calculate the number of appearance of $w_i$ for each text tonality class separately
   $$x_{iz} = \sum_{i \in W, j \in T} w_{ij}, z = \frac{1}{3}.$$
   2. Calculate number of appearance of $s_z$ in training template $T$
   $$y_z = \sum_{j \in T} s_{zj}, z = \frac{1}{3}.$$
   3. Calculate conditional probability $P(W_i|S_z)$ of occurrence $w_i$ in $z$-th text tonality class
4. Calculate probability \( P(s_x) \) of \( T_y \)-th text assignment to the \( z \)-th text tonality class

\[
P(s_x) = \frac{y_x}{\sum_{x=1}^{3}y_x}.
\]

II. Using of the Naïve Bayesian Classifier:

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

\[
P\left( \frac{s_z}{\{w_{ij+1}\}} \right) = P(s_z) \times \prod_{x} P\left( \frac{w_x}{s_z} \right).
\]

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

\[
R_z = P\left( \frac{s_z}{\{w_{ij+1}\}} \right),
\]

\[
s_{z+1} = \arg \max_{z=1}^{3} R_z.
\]

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 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 \( 1 - of - N \) 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 \( s(\cdot-1) \). 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 \( 1 - of - N \) 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 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, uniqueness; 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 I.

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](image)

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.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
F – \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

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.

| 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://eurekacengine.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.
References (transliterated)

1. Ameru 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://eurekacgnine.ru/description (accessed 15.09.2019).

3. Huang M., Zhang 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/s1049-9N019–05783-5.

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. Volume 2. P. 60–70. DOI:10.15588/1607-3274-2019-2-7. e-ISSN 1607-3274.

6. Mikolov T., Karafiát M., Burget L., Černocky 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: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 13th 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 Couriers». URL: http://www.iteco.ru/solutions/business_intelligence_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., Ніфілін В. В. Автоматизоване створення електронного словника. Інформаційні технології: наука, техніка, технологія, освіта, здоров’я: тези доповідей XXV Міжнародної науково-практичної конференції МікроCAD. Ч. 1. Харків: НТУ «ХПІ», 2017. С. 32.

17. Борисова Н. В., Ніфілін В. В. Застосування методів корпусної лінгвістики для дослідження особливостей використання сучасного молідзінного словника. Інформаційні технології: наука, техніка, технологія, освіта, здоров’я: тези доповідей XXVI Міжнародної науково-практичної конференції МікроCAD. Ч. 1. Харків: НТУ «ХПІ», 2018. С. 27.

18. Корпус коротких текстов RuTweetCorp. URL: http://study.mokoron.com (access date: 15.09.2019).

19. Romanov A. V., Vasilevich M. I., Kurtkova A. V., Mezerykov R. V. Analiz tonalnosti tekstov s ispolzovaniem metodov mashinnoj obuchenija [Sentiment Analysis of Text Using Machine Learning Techniques]. 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. Створення словника електронного словника на основі автоматизованої використання електронного словника. Інформаційні технології: наука, техніка, технологія, освіта, здоров’я: тези доповідей XXV Міжнародної науково-практичної конференції MicroCAD-2017. Ч. 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.

21. Borysova N. V., Nifilin V. V. Zastosuvannia metodiv korpusnoi lingvistiki dlia doslidzhennia osoblyvostei vkyorkorstannia suchasnoho molidzinha sljenga [Using of corpus linguistics methods to study the features of modern youth slang]. Informacnyi technologi: nauka, technika, technologiya, osvita, zdorov’ja: tez dopovidei XXV Mjachnorochnoi naukovo-praktchnoi konferenci 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.

22. Korpus korotkih tekstov RuTweetCorp [Short texts corpus RuTweetCorp]. Available at: http://study.mokoron.com (accessed 15.09.2019).

23. Romanov A. V., Vasilevich M. I., Kurtkova A. V., Meshcheriakov R. V. Analiz tonalnosti tekstov s ispolzovaniem metodov mashinnoj obuchenija [Sentiment Analysis of Text Using Machine Learning Techniques]. Proceedings of the R. Piotrowski’s Readings in Language Engineering and Applied Linguistics. CEUR Workshop Proceedings. Volume 2233. Saint Petersburg, Russia, 2017, pp. 86–95.
Відомості про авторів / Сведения об авторах / About the Authors

Борисова Наталія Володимирівна – 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 (Melynk 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

ВОГРУНТУВАННЯ 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 спосіб 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, обсяг ресурсів, вартість обробки та інші. Вихідними даними для прийняття рішень по вибору крашою архітектури є характеристики даних задач, алгоритми обробки, характеристики прийнятних типів архітектури обчислювальних підрозділів, умови і вимоги до організації обчислювальних процесів та процесів управління, процедур обробки, їх характеристики і параметри, особливості програмного середовища, інструментальних засобів розробки і модифікації програмних рішень. Наявність неизніченості, викликаної майбутньою аспекти 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 робочому навантаженні, яке змінюється у часі, що проявляється як у усуненні задач обробки та їх вихідних даних, так і в необхідності процедурах обробки. Це 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 рівня їх узгодження. Варіанти архітектури системи, що відповідають узгодженим рішенням, складають підмножину, яка надає обґрунтований варіант вибору рішень, які 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 числових розрахунків та аналізу отриманих результатів.

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