

Method of Determining the Text Sentiment by Thematic Rubrics

Anatoliy Sachenko^{1,2}, Taras Lendiuk¹, Khrystyna Lipianina-Honcharenko¹, Maciej Dobrowolski², Gena Boguta¹ and Leonid Bytsyura¹

¹ West Ukrainian National University, Lvivska str., 11, Ternopil, 46000, Ukraine

² Kazimierz Pulaski University of Technology and Humanities in Radom, Department of Informatics, Jacek Malczewski str., 29, Radom, 26 600, Poland

Abstract

A method for determining the text sentiment of by thematic rubrics is proposed. It is based on a complex approach that integrates natural language processing, machine learning and linguistic analysis for automatic classification of text data. To implement the method, a generalized client-server architecture of the text sentiment for analysis system was developed and a set of data was collected from a wide range of article texts from the Internet sites of Ukraine, which ensure the representativeness of various styles, genres and topics. The high efficiency of the system in terms of classification of texts by rubrics was experimentally confirmed with a correspondence of 92%.

Keywords

text sentiment, thematic rubrics, natural language processing, machine learning, linguistic analysis, automatic classification, text data

1. Introduction

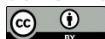
Currently, we are witnessing many information wars being waged on the world stage, with a special emphasis on the war in Ukraine. Russia uses information operations as a key element of its hybrid war against Ukraine and directs significant resources to the creation and dissemination of disinformation, the purpose of which is to influence public opinion, undermine trust in the Ukrainian state, its institutions and leaders, as well as to distort the real state of affairs in the international community

In the context of increasing information attacks from Russia, the importance of an accurate and objective analysis of the text sentiment of related to various aspects of the war and its impact on various spheres of life becomes extremely relevant. Automated analysis of the emotional coloring of texts can help detect attempts to manipulate public opinion, assess the general mood in society regarding certain events

¹COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

✉ as@wunu.edu.ua (A. Sachenko); tl@wunu.edu.ua (T. Lendiuk); xrustya.com@gmail.com (Kh. Lipianina-Honcharenko); m.dobrowolski@uthrad.pl (M. Dobrowolski); genaboguta7@gmail.com (G. Boguta); l.bytsyura@wunu.edu.ua (L. Bytsyura)

ORCID 0000-0002-0907-3682 (A. Sachenko); 0000-0001-9484-8333 (T. Lendiuk); 0000-0002-2441-6292 (Kh. Lipianina-Honcharenko); 0000-0003-0296-9651 (M. Dobrowolski); 0009-0000-9788-1753 (G. Boguta); 0000-0002-9476-011X (L. Bytsyura)



© 2024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

or policies, and identify changes in the information space that may indicate new directions of information attacks or disinformation campaigns.

The proposed method integrates advanced technologies of natural language processing (NLP), machine learning and linguistic analysis, which allows not only to automate the process of determining the emotional coloring of texts, but also to ensure high accuracy and objectivity of the obtained results. The application of such a complex approach in the conditions of information warfare opens up new opportunities for monitoring the information space, identifying and analyzing information operations, as well as for developing effective strategies for countering disinformation. Given the high dynamics of the information space and the constant change in the tactics and strategies of information warfare, the development and implementation of innovative text analysis methods that allow for prompt monitoring and analysis of attempts at manipulation and disinformation is extremely important.

Further materials are presented according to the following structure. In Chapter 2 the analysis of recent related works is fulfilled and in Chapter 3, the proposed method is described. Chapter 4 is dedicated to implementation and case studies, and Chapter 5 summarizes the received outcomes.

2. Related Work

The study of text tonality, which involves the integration of natural language processing (NLP), machine learning, and linguistic analysis, occupies an important place in modern scientific research. In [1], the analysis tool for text sentiment was developed based on the fusion of machine learning algorithms, demonstrating the effectiveness of combining different methods to more accurately determine the emotional coloring of texts. Authors [2] emphasized the importance of using TF-IDF and machine learning methods for sentiment analysis of Twitter data, confirming the importance of these techniques in the field of big data processing. In [3], the impact of the lexical richness of the training corpus on the performance of machine learning in analysis tasks was highlighted, emphasizing the need for a careful approach to data selection and processing. Authors [4] presented an innovative approach to the classification of hierarchical comments using BERT and a specialized naive Bayes classifier, opening up new opportunities for advanced text analysis. In [5], a deep learning-based approach to sentiment analysis is proposed for ranking online products using a set of probabilistic linguistic terms, which demonstrates the potential of using these technologies for commercial purposes. Authors [6] developed a technique for analyzing the sentiment of data from Twitter using NLP and machine learning, which emphasizes the importance of an integrated approach to information processing. The authors [7] demonstrate how sentiment analysis on Twitter can be improved using NLP and machine learning methods for emotion categorization and trend visualization. In [8], the advantages of integrating linguistic analysis for the study of human language are revealed, which opens up new opportunities for accurate thematic categorization. Research [9] focuses on the analysis of online product reviews using advanced deep learning and machine learning techniques [23, 24] to improve data classification and extract detailed emotion information [21].

Compared to the analogue [10], which focuses on the sentiment analysis of Twitter data, this work is distinguished by a deep integration of natural language processing (NLP) techniques, machine learning and linguistic analysis. In addition, a distinctive feature of the proposed approach is the development of specialized algorithms for accurate determination of sentiment taking into account contextual semantics and lexical diversity within specific thematic headings. This allows not only to more accurately classify the emotional coloring of texts, but also to detect subtle nuances in emotional expressions, which makes the proposed approach more adaptable to the complexities of natural language and provides a higher accuracy of analysis compared to existing methods.

3. Proposed Method

The proposed method for determining the sentiment of text by thematic headings is a comprehensive approach that integrates natural language processing (NLP), machine learning, and linguistic analysis for automatic classification of text data. The method allows to evaluate the emotional color of the text (positive, neutral or negative) and to express it quantitatively in the form of an expanded percentage scale in the range from -100% to +100%. Let us present this method as a set of the following stages and structurally (Fig. 1):

Stage 1. Text pre-processing [11]:

- 1.1. Removal of extra characters, text normalization;
- 1.2. Detection and removal of stop words;
- 1.3. Tokenization of text [11].

Stage 2. Classification of the text by thematic headings [12]:

- 2.1. The use of machine learning methods to determine the thematic rubric of the text [13];
- 2.2. Training models on a predefined set of texts with a clear thematic affiliation.

Stage 3. Creation and use of dictionaries for each thematic rubric [14]:

- 3.1. Development of dictionaries with key words and phrases specific to each rubric;
- 3.2. Determination of the emotional coloring of keywords (positive, negative, neutral).

Stage 4. Sentiment analysis [15]:

- 4.1. Using sentiment computation methods such as dictionary-based estimation or deep learning models;
- 4.2. Calculation of the sentiment index T for the text based on the formula:

$$T = \frac{P - N}{P + N + Q} \times 100\%$$

where P is the number of positive words, N is the number of negative words, Q is the number of neutral words.

Stage 5. Inversion of tonality. In the case when the text belongs to a hostile source and does not contain a direct mention of Ukraine, tonality inversion is used.

Stage 6. Displaying the results. Development of an interface for visualizing the tonality of the text with the possibility of viewing a detailed analysis by thematic headings.

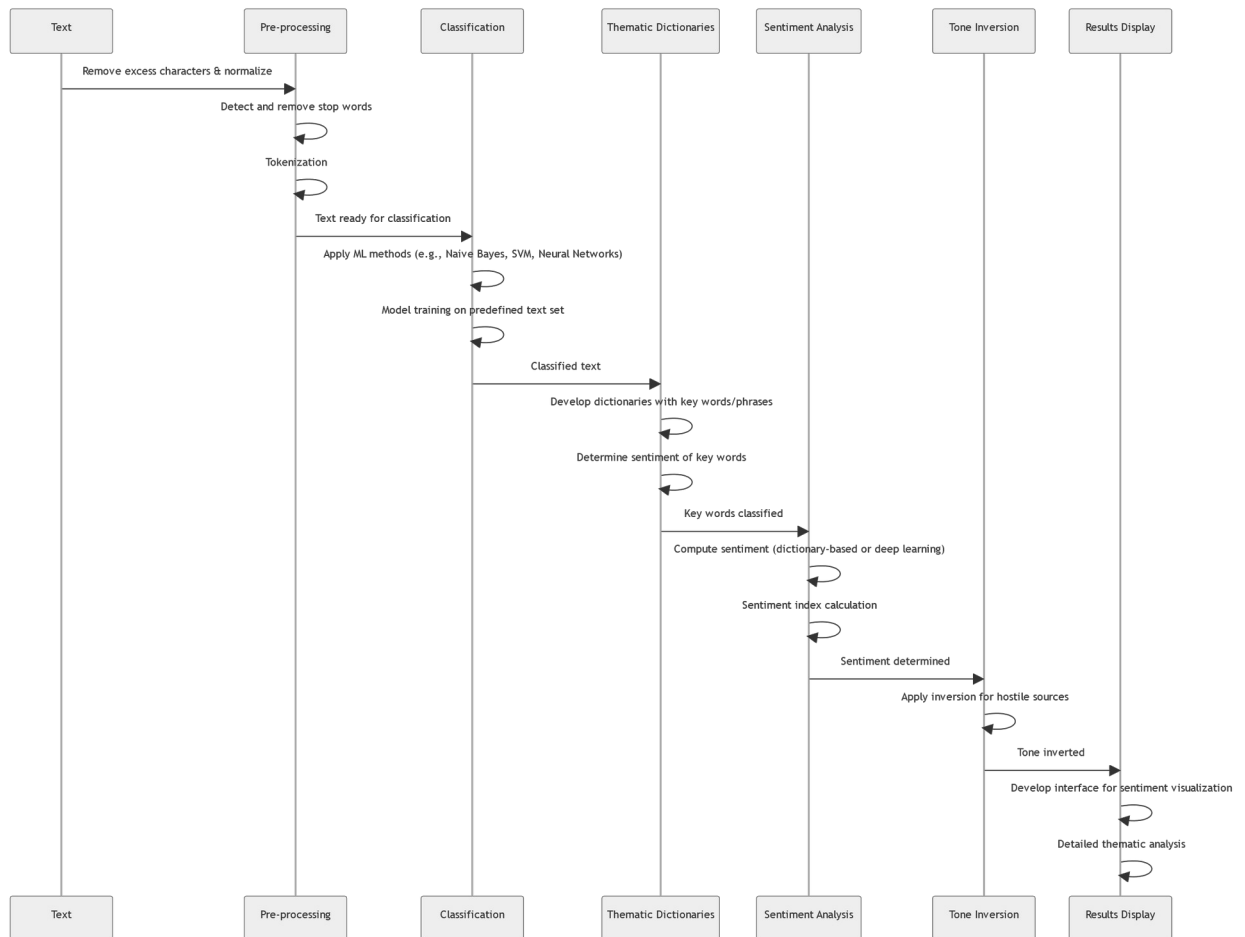


Figure 1: Structure of the method of determining the tonality of the text by thematic rubrics

4. Implementation and Case Study

The generalized client-server architecture of the text tonality analysis system is presented in Fig. 2. The client initiates the interaction by sending a request to the server, which in turn processes the received information. After processing, the server sends the necessary information back to the client. This process is cyclic, so after transferring information, the client can initiate interaction with the server again. The diagram shows a typical request-response model that is fundamental to a client-server architecture, where the server acts as a provider of resources and services and the client acts as a consumer of those resources and services.

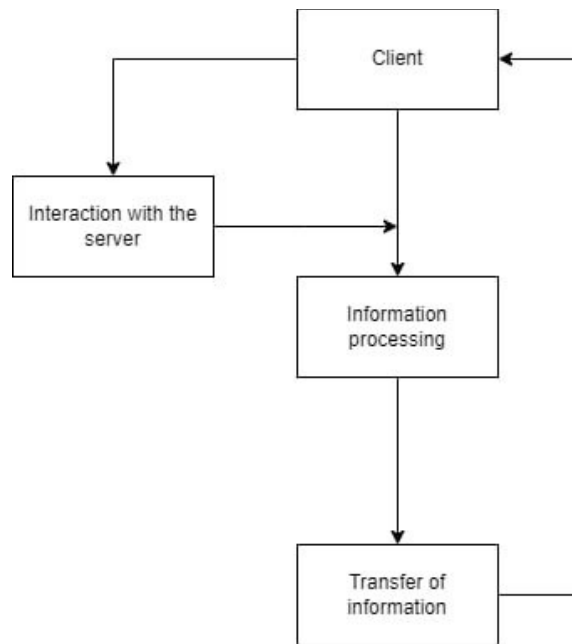


Figure 2: Generalized client-server architecture of the text tonality analysis system

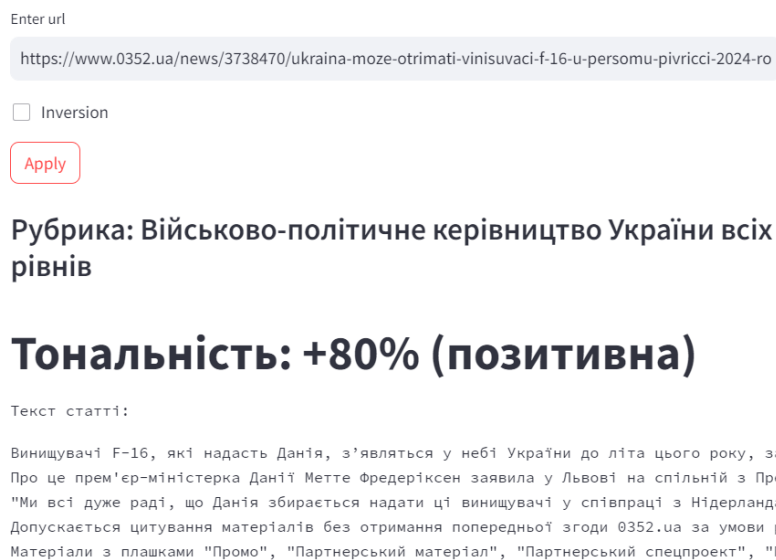
To implement the proposed method, a set of data was collected from a wide range of texts of articles from Internet sites of Ukraine, which ensure the representativeness of various styles, genres and topics. The sample covers a variety of emotional contexts to test the algorithm's ability to accurately classify emotional nuances. The analysis of the text will be assigned to one of the following thematic headings, which were determined by experts from the security service of state bodies:

- Military and political leadership of Ukraine at all levels;
- Law enforcement agencies of Ukraine;
- Armed Forces of Ukraine;
- Socio-political situation in the regions of Ukraine (attitudes towards mobilization, socio-economic stability, etc.);
- Pro-Russian religious organizations on the territory of Ukraine;
- Pro-Russian movements, formation of the concept of "Russian peace";
- International image of Ukraine in the EU (English, German, Polish, Romanian, French, Hungarian, Ukrainian, Russian languages);
- International image of Ukraine in the USA, Canada and Great Britain (English language);
- International image of Ukraine in African countries (English, French, Arabic languages);
- International image of Ukraine in Asian countries (Chinese, Russian, Turkish, Arabic, Georgian, Kazakh, Farsi, Kyrgyz, Tajik, Uzbek, Japanese, Korean languages);
- Ukraine in the information space of the Russian Federation;
- Ukraine in the information space of the Republic of Belarus;

- Socio-economic, political, military situation in the Russian Federation (attitudes towards the top management, mobilization, deterioration of the economic situation, etc.).

The main functionality of the web interface (Fig. 3) [15] allows users to insert url links of news sites for analysis. After submitting the text, the system uses an algorithm to determine the emotional color, displaying the results in the form of an understandable report. The report includes quantitative indicators of the presence of positive and negative words, as well as an overall text tone index.

As a result of the analysis of the tonality of the text [16], the rubric "Military and political leadership of Ukraine at all levels" was determined, where a positive tone of +80% was indicated. This reflects a high level of positive emotional coloring of the text of the article. Details are given about military aspects, including the acquisition or use of military equipment (the F-16 and Bayraktar are mentioned), and an optimistic attitude towards events related to the war in Ukraine is reflected.



Enter url

<https://www.0352.ua/news/3738470/ukraina-moze-otrimati-vinisuvaci-f-16-u-persomu-pivricci-2024-ro>

Inversion

Apply

Рубрика: Військово-політичне керівництво України всіх рівнів

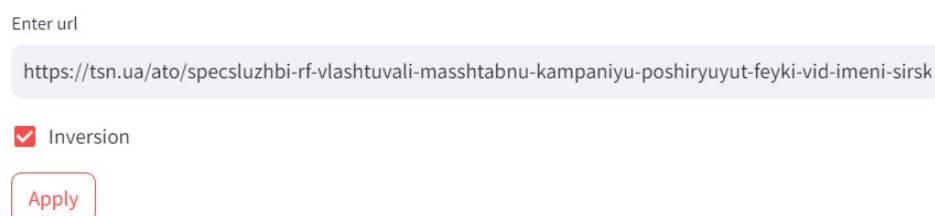
Тональність: +80% (позитивна)

Текст статті:

Винищувачі F-16, які надасть Данія, з'являться у небі України до літа цього року, за Про це прем'єр-міністерка Данії Метте Фредеріксен заявила у Львові на спільній з Пре "Ми всі дуже раді, що Данія збирається надати ці винищувачі у співпраці з Нідерланда Допускається цитування матеріалів без отримання попередньої згоди 0352.ua за умови р Матеріали з плашками "Промо", "Партнерський матеріал", "Партнерський спецпроект", "П

Figure 3: Example of text analysis result

The inversion function (Fig. 4) of the results was included for users who wish to analyze the opposite emotional tone of the text. This can be useful, for example, when analyzing texts containing sarcasm or irony, where the literal meaning of the words can be misleading. The inversion allows you to quickly see how the overall assessment of emotional coloring will change if these stylistic figures are taken into account.



Enter url

<https://tsn.ua/ato/specsluzhbi-rf-vlashtuvali-masshtabnu-kampaniyu-poshiryuyut-feyki-vid-imeni-sirsk>

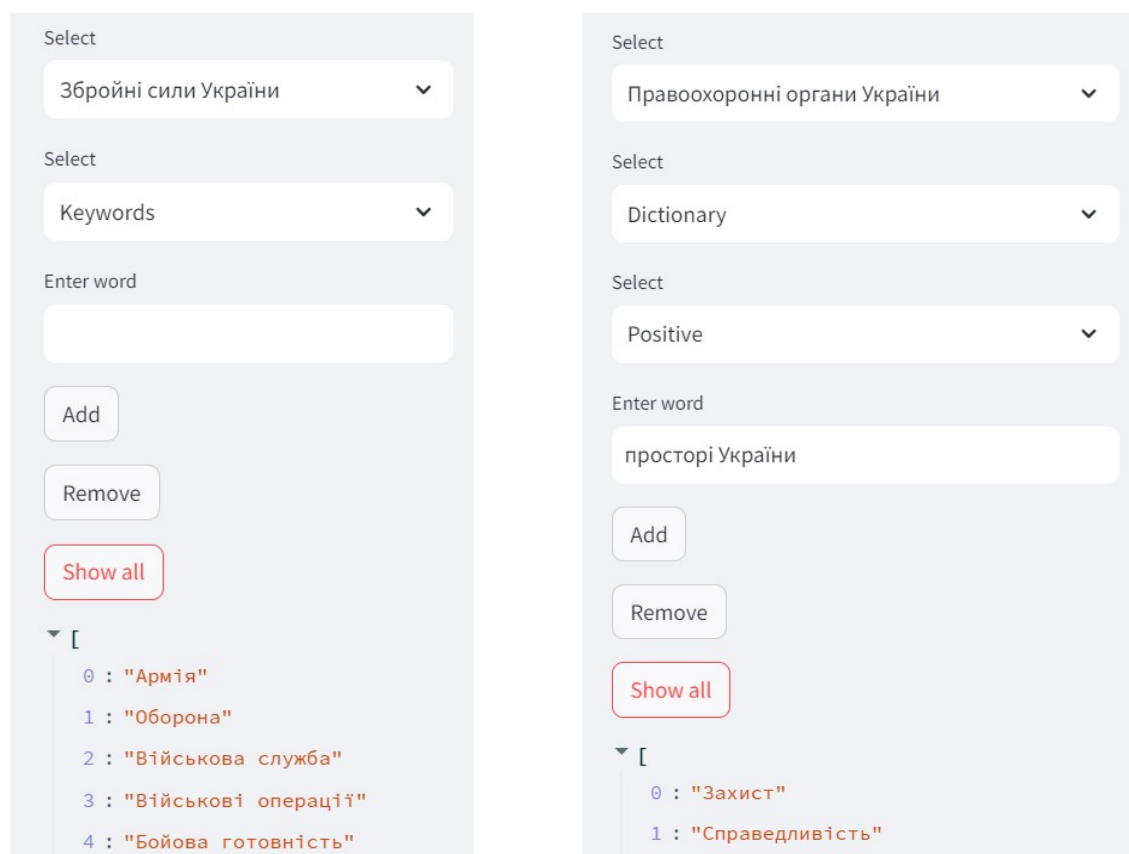
Inversion

Apply

Figure 4: Field for entering a link and using inversion

One of the most important aspects of the web interface (Fig. 5a) is the ability to edit keywords for each heading. Users can add new words to dictionaries, remove existing ones, which affects their significance in the analysis algorithm. This allows you to adapt the system to the specific needs of the user and increase the accuracy of determining the tonality for specific topics or writing styles.

In addition, the web interface includes a dictionary management module (Fig. 5c), which allows users to view and update the database of words on the basis of which the analysis is performed. This is especially important to take into account linguistic updates, social and cultural changes that can affect language and its emotional load.



a – key words

b – dictionary of emotions

Figure 5: Example of editing keywords and dictionary

A comparative analysis was conducted to assess the effectiveness of the system for determining the tonality of texts (100 news stories) by comparing the automated results of the system with experts' assessments. The analysis took into account such parameters as the assignment of the text to the appropriate rubric by the system and experts, the comparison of the tonality of the texts determined by both the system and experts, and the correspondence of these assessments. Using the URL as a unique identifier for each text ensured accurate tracking of results. In addition to the quantitative evaluations, the experts provided additional notes for a deeper understanding of the reasons for the discrepancies between the expert evaluations and the results of the system, which will contribute to the further improvement of its accuracy and reliability.

Based on the results of filling out Table 1, statistical indicators were calculated, namely the simple percentage of matches between the system and experts.

Table 1
Evaluation of System Efficiency

Statistical indicator	indicator value
Correspondence of rubrics	92%
Average tonality deviation	15%

The obtained results (see Table 1) indicate a fairly high efficiency of the system in terms of classification of texts by headings with a correspondence of 92%. This means that in most cases the system correctly identifies the thematic category of the text, which indicates its reliability in determining the context of news. However, the average tonality deviation of 15% is quite significant and may indicate some shortcomings in the work of the algorithms of the system for evaluating the emotional coloring of the text. This may be due to the incompleteness of the dictionaries used for sentiment analysis. However, dictionaries can be constantly updated, which is a significant advantage of the system, since the language is constantly evolving, and the context and use of the vocabulary can change. The ability to supplement dictionaries by users allows the system to adapt more quickly to novelties in the language and changes in the use of terms, especially in the field of news, where it is important to take into account not only the lexical meaning of words, but also their connotative influence.

Thus, the system demonstrates a high accuracy in the classification of thematic headings, but needs improvement in determining tonality. Constant addition and updating of dictionaries, with the possibility of making changes from users, is an important process for increasing the accuracy of the analysis of the emotional coloring of texts.

5. Conclusion

A method for determining the text sentiment by thematic rubrics is proposed based on an integrated approach that integrates natural language processing, machine learning and linguistic analysis for automatic classification of text data. This allows you to assess the emotional color of the text (positive, neutral or negative) and express it quantitatively in the form of a percentage scale from -100% to +100%.

The text sentiment analysis system implemented based on the method has a high accuracy (92%) of classification of thematic headings. This means that in most cases the system correctly identifies the thematic category of the text, which indicates its reliability in determining the context of news.

However, the average sentiment deviation of 15% is quite significant and may indicate some shortcomings in the work of the algorithms of the system for evaluating

the emotional coloring of the text. This may be due to the incompleteness of the dictionaries used for sentiment analysis.

In the future, authors are going to explore the methods [17-20] for improving the quality and performance of text sentiment analysis.

References

- [1] P. Ajitha, A. Sivasangari, R. Rajkumar, & S. Poonguzhali, Design of text sentiment analysis tool using feature extraction based on fusing machine learning algorithms. *Journal of Intelligent & Fuzzy Systems* 40 (2021), 6375-6383 <https://dx.doi.org/10.3233/jifs-189478>
- [2] S. Singh, K. Kumar, & B. Kumar, Sentiment analysis of twitter data using TF-IDF and machine learning techniques, in: *Proceedings of the 2022 IEEE 6th International Conference on Communication and Electronics Systems (ICCES)*, 2022, pp. 252-255. <https://dx.doi.org/10.1109/com-it-con54601.2022.9850477>
- [3] S. Garg, A. Saini, & N. Khanna, Is sentiment analysis an art or a science? Impact of lexical richness in training corpus on machine learning. in: *Proceedings of the 2016 IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2729-2735. <https://dx.doi.org/10.1109/ICACCI.2016.7732474>
- [4] M. Dhina, & S. Sumathi, An innovative approach to classify hierarchical remarks with multi-class using BERT and customized naïve bayes classifier. *International Journal of Engineering, Science and Technology*, 13 (2022), 32-45. <https://dx.doi.org/10.4314/ijest.v13i4.4>
- [5] Z. Liu, H. Liao, M. Li, Q. Yang, & F. A Meng, Deep learning-based sentiment analysis approach for online product ranking with probabilistic linguistic term sets, *IEEE Transactions on Engineering Management* (2023). <https://dx.doi.org/10.1109/tem.2023.3271597>
- [6] K. Brindha, S. Senthilkumar, A. K. Singh, & P. M. Sharma, Sentiment analysis with NLP on Twitter data, in: *Proceedings of the IEEE International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2022, pp. 1-5. <https://dx.doi.org/10.1109/SMARTGENCON56628.2022.10084036>
- [7] K. Darshan, J. Samuel, M. Swamy, P. Koparde, & N. Shivashankara, NLP - Powered sentiment analysis on the Twitter, *Saudi Journal of Engineering and Technology* 9, (2024) 1-11. <https://dx.doi.org/10.36348/sjet.2024.v09i01.001>
- [8] P. Srinivas, K. Gayathri, K. Bhavitha, Jahnavi & K. D. Sarath, BLIP-NLP model for sentiment analysis, in: *Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, 2023, pp. 468-475. IEEE <https://dx.doi.org/10.1109/ICECAA58104.2023.10212253>
- [9] L. Bharadwaj, Sentiment analysis in online product reviews: mining customer opinions for sentiment classification 5, (2023). <https://dx.doi.org/10.36948/ijfmr.2023.v05i05.6090>

- [10] S. Voloshyn, V. Vysotska, O. Markiv, I. Dyyak, I. Budz and V. Schuchmann, Sentiment analysis technology of English newspapers quotes based on neural network as public opinion influences identification tool, in: Proceedings of the 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 83-88, doi: 10.1109/CSIT56902.2022.10000627.
- [11] R. Gramyak, H. Lipyana-Goncharenko, A. Sachenko, T. Lendyuk, & D. Zahorodnia, Intelligent Method of a competitive product choosing based on the emotional feedbacks coloring, in: Proceedings of the IntelITSIS, 2021, pp. 246-257.
- [12] H. Lipyana, S. Sachenko, T. Lendyuk, & A. Sachenko, Targeting model of HEI video marketing based on classification tree, in: Proceedings of the ICTERI Workshops, October 2020, pp. 487-498.
- [13] H. Lipyana, S. Sachenko, T. Lendyuk, V. Brych, V. Yatskiv, & O. Osolinskiy, Method of detecting a fictitious company on the machine learning base, in: International Conference on Computer Science, Engineering and Education Applications, January, 2021, pp. 138-146. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-80472-5_12
- [14] K. Lipianina-Honcharenko, T. Lendiuk, A. Sachenko, O. Osolinskiyi, D. Zahorodnia, & M. Komar, An intelligent method for forming the advertising content of higher education institutions based on semantic analysis, in: International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications, September 2021, pp. 169-182. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-14841-5_11
- [15] Streamlit. URL: <https://nelczgkwcsghtwdkw7buxn.streamlit.app/>
- [16] Ukraine can receive F-16 fighter jets in the first half of 2024, - the premier of Denmark, 2024. URL: <https://www.0352.ua/news/3738470/ukraina-moze-otrimati-vinisuvaci-f-16-u-persomu-pivricci-2024-roku-premerka-danii>
- [17] M. Komar, V. Golovko, A. Sachenko and S. Bezobrazov, Development of neural network immune detectors for computer attacks recognition and classification, in: Proceedings of the 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), Berlin, Germany, 2013, pp. 665-668, doi: 10.1109/IDAACS.2013.6663008.
- [18] Zhengbing Hu, Yevgeniy V. Bodyanskiy, Nonna Ye. Kulishova, Oleksii K. Tyshchenko, A multidimensional extended neo-fuzzy neuron for facial expression recognition, International Journal of Intelligent Systems and Applications (IJISA) 9 (2017) 29-36. DOI: 10.5815/ijisa.2017.09.04
- [19] V. Vysotska, O. Markiv, S. Tchynetskyi, B. Polishchuk, O. Bratasyuk, V. Panasyuk, Sentiment analysis of information space as feedback of target audience for regional e-business support in Ukraine, in: CEUR Workshop Proceedings, vol-3426, 2023, 488-513.
- [20] Sutriawan, S., P. N. Andono, M. Muljono, & R. A. Pramunendar, Performance evaluation of classification algorithm for movie review sentiment analysis,

- [21] S. Voloshyn, O. Markiv, V. Vysotska, I. Dyyak, L. Chyrun and V. Panasyuk, Emotion recognition system project of English newspapers to regional e-business adaptation, in: Proceedings of the 2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 392-397, doi: 10.1109/CSIT56902.2022.10000527.
- [22] K. Mehta, & S. P. Panda, Sentiment analysis on e-commerce apparels using convolutional neural network, International Journal of Computing, vol. 21, issue 2, pp. 234-241, 2022. <https://doi.org/10.47839/ijc.21.2.2592>
- [23] R. Lynnyk, V. Vysotska, Y. Matseliukh, Y. Burov, L. Demkiv, A. Zaverbnyj, A. Sachenko, I. Shylinska, I. Yevseyeva, and O. Bihun, DDOS attacks analysis based on machine learning in challenges of global changes, in: 2020 CEUR Workshop Proceedings, 2020 vol. 2631, pp. 159-171.
- [24] O. Soprun, M. Bublyk, Y. Matseliukh, V. Andrunyk, L. Chyrun, I. Dyyak, A. Yakovlev, M. Emmerich, O. Osolinsky, A. Sachenko, Forecasting temperatures of a synchronous motor with permanent magnets using machine learning, in: 2020 CEUR Workshop Proceedings, 2020, vol. 2631, pp. 95-120.