Issue 15, 2024

https://doi.org/10.23939/sisn2024.15.074

UDC 004.415

METHODS OF MACHINE LEARNING AND DESIGN OF A SYSTEM FOR DETERMINING THE EMOTIONAL COLORING OF UKRAINIAN-LANGUAGE CONTENT

Anton Lomovatskyi[1], Taras Basyuk[2]

Lviv Polytechnic National University,
Department of Information Systems and Networks, Lviv, Ukraine
[1] E-mail: anton.a.lomovatskyi@lpnu.ua, ORCID: 0009-0004-5170-3272
[2] E-mail: Taras.M.Basyuk@lpnu.ua, ORCID: 0000-0003-0813-0785

In the article, the authors analyze the current state of research in the field of emotional analysis of Ukrainian-language content for data mining systems. The main methods and approaches to solving the problem are analyzed. The main machine learning algorithms for analyzing textual content are also considered. As a result of the analysis, the main methods and approaches that can be used to analyze the Ukrainian language were identified and classified. The next step was to design the system's functionality using a structural approach. The authors of the article have designed an information system using a structural approach. A contextual diagram of the information system was developed and its main process was decomposed in order to show in more detail the process of preparing and analyzing information in the process of determining the emotional coloring.

Key words: emotional analysis; machine learning; designing; classification of methods; information technology.

Introduction. Formulation of the problem

The emotional coloring of content is determined by the emotions and feelings experienced by the consumer or audience. These can be negative, positive, or neutral emotions conveyed through phrases, words, and context. Emotional coloring is used in various fields of human activity, such as marketing, advertising, social media, and customer feedback analysis [1].

The Ukrainian language is multifaceted and variable in the meaning of words (the same word can have different meanings depending on the context and emphasis). That is why the development of natural language processing systems that take into account the emotional coloring and peculiarities of the Ukrainian language is key to improving the quality and efficiency of Ukrainian-language content development [2].

Several key theses can be identified to define the areas of analysis of the state of research on the emotional coloring of content [3]:

- The importance of understanding emotional coloring. Understanding this allows us to better choose the tone of the content, which will allow the reader to better understand the message.
- Analysis of user feedback. Understanding this will improve the quality of service.
- Emotional analytics in social media. Understanding this allows us to improve marketing strategies and reputation management systems.

- Psychological analysis of the text. Understanding this allows us to determine the contingent of different groups of people.

Emotional analysis of content is an extremely important component of the development of Internet technologies and modern communication systems. Understanding the context is an extremely powerful tool for automating and optimizing processes.

The connection of the highlighted problem with important scientific and practical tasks

Problems in the emotional analysis of Ukrainian-language content include:

- Language features. The Ukrainian language is linguistically complex with a variety of rules and exceptions.
- Insufficient data. Insufficient training data base can affect the accuracy of the analysis result.
- Semantic complexity. Words and phrases can have different meanings depending on the context.
- Polarized emotions. The emotional coloring of a sentence can be complex, i.e. contain combinations of different emotions.

In view of this, an important scientific and practical task is to classify machine learning methods for determining the emotional color of Ukrainian-language content for data mining systems. The solution to this task will provide the necessary tools for creating a system for determining the emotional color of Ukrainian-language content.

Formulation of the purpose of the article

The purpose of the study is to analyze the main methods of determining the emotional coloring of Ukrainian-language content and to design a system. The study will provide a means of systematizing and comparing different approaches, which will help to expand understanding in the field of analyzing the emotional component of texts and the development of relevant technologies. To achieve this goal, it is necessary to solve the following main tasks: to analyze the known methods of emotional analysis of texts; to identify the main factors that affect the determination of emotional coloration; to classify machine learning methods for emotional analysis; design the system using a structural approach.

The results of the work performed in their entirety solve the urgent scientific and practical task of developing effective methods of emotional analysis of Ukrainian-language content and serve as the basis for creating methodological support in this subject area.

Analysis of existing methods

The analysis of research has shown that there are many approaches to solving this problem, among the most common are: Lexicon-based Approaches, Rule-based Approaches, Word Embeddings, Supervised Learning. A combination or individual use of each of these methods will help to develop a system for determining the emotional coloring of Ukrainian-language content.

Lexicon-based approaches

Lexicon-based approaches – the direction of text analysis and emotional coloring is based on the use of a certain vocabulary or vocabularies that are associated with real emotions. These approaches allow us to determine the polarity of a word or sentence. The algorithm of such approaches is quite simple [4]:

1. Creating or using an existing lexicon (dictionary) that contains an assessment of various emotionally colored words.
2. Counting emotional words in the text. In general, this is the number of positive, negative and neutral words.
3. Assigning weights to words. Because certain words can occur several times and have different strengths of emotional association. This step allows us to determine the final result even more accurately.
4. Summarize the scores of emotional words. This can be a simple count with the addition of different types of words, or more complex statistical methods using algorithms and weights.

One of the most famous existing dictionaries is EmoLex. EmoLex (NRC Emotion Lexicon) is an example of a dictionary designed specifically for detecting emotions in text. It is one of the most popular and widely used emotion dictionaries in the field of natural language processing (NLP). This lexicon was developed by the National Research Council of Canada for use in natural language processing (NLP).

The article [5] describes the use of dictionary-based methods, especially the EmoLex dictionary, to analyze social sentiment during the 2019 presidential election in Nigeria. The authors analyzed more than 220 thousand tweets and grouped them by emotions (sadness, trust, anger, fear, joy, expectation, disgust, surprise) and by mood (positive, negative). To achieve this result, the authors created the following architecture (Fig. 1).

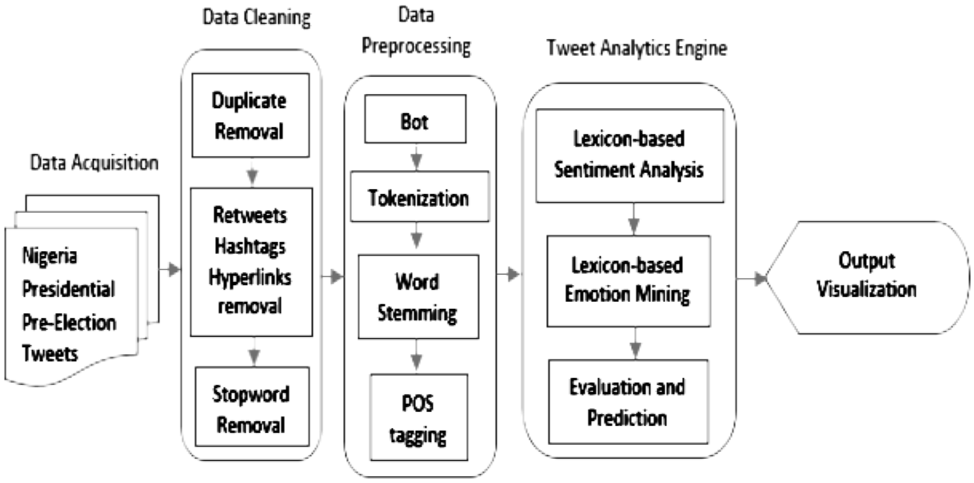

Fig. 1. Architecture of the system for analyzing social sentiment

That is, before the emotion analysis itself, the text needs to be cleaned (duplicates removed) and processed (tokenization, bringing words back to their original form). Then, using R methods and the NRC Word Emotion Association Lexicon (EmoLex), these words were assigned scores based on Robert Plutchik's emotional scores. The results of these analyses were then compared in graphs and tables.

In this article [6], the authors analyzed more than 35 million reviews of various video games of 11 different genres on the Steam platform. They also divided them into 8 emotions and 2 moods. The authors very successfully identified irrelevant reviews, and this is very important in the process of data scraping. Thus, irrelevant reviews were those that users find funny when the player received the game for free, when the player has only one game, if the review was written during the early access period of the game. The authors used the syuzhet utility for R, namely the get_nrc_sentiment function. In conclusion, the authors noted that negative reviews contain more information about the user's experience, but the average length of a review is only slightly longer than the length of a positive review.

For the Ukrainian language, us can use this dictionary (Table 1), as it officially supports the Ukrainian localization. It can be downloaded by following the link to the library's official website.

Table 1

Using the NRC EmoLex dictionary for the Ukrainian language

English Word	anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	Ukrainian Word
aback	0	0	0	0	0	0	0	0	0	назад
abacus	0	0	0	0	0	0	0	0	0	рахівниця
abandon	0	0	0	1	0	1	0	1	0	відмовитися
abandoned	1	0	0	1	0	1	0	1	0	покинутий
abandonment	1	0	0	1	0	1	0	1	1	занедбаність
abate	0	0	0	0	0	0	0	0	0	зменшуватися
abatement	0	0	0	0	0	0	0	0	0	зменшення

Rule-based approaches

Rule-based approaches, in the context of text analysis and emotion detection, are based on predefined rules or rules that are programmatically set to detect specific emotional features in a text. These approaches use a set of rules or templates to identify emotional expressions, phrases, or contexts [7].

TextBlob is a text processing library in the Python programming language that provides sentiment analysis and emotional coloration detection capabilities. While it is commonly used for sentiment analysis, it can also be extended to recognize emotional features in text.

In this article [8], the authors explore social attitudes toward COVID-19 by analyzing 2 million tweets. The general principle of the system is shown in the diagram (Fig. 2):

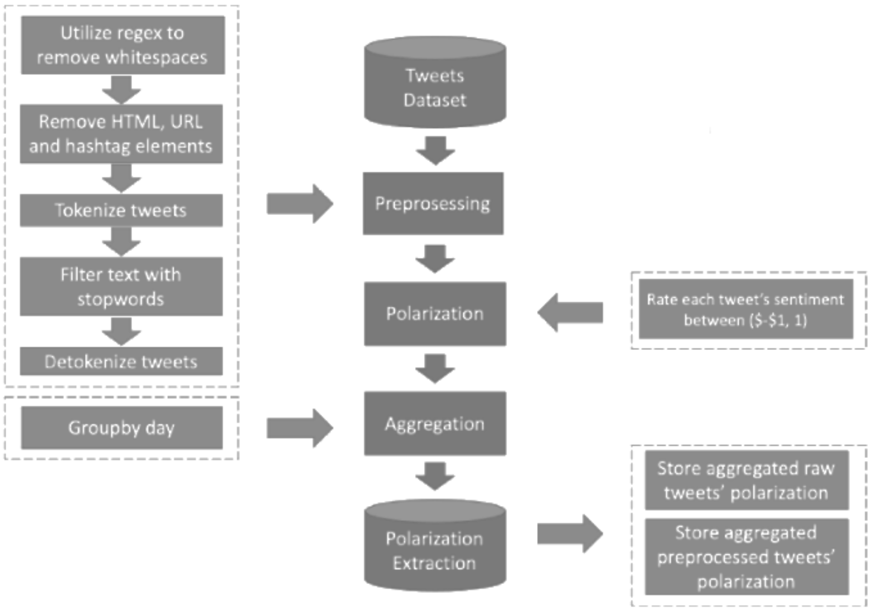

Fig. 2. Scheme of the system for analyzing tweets

It is important to note the stages of analysis:

1. Creating a dataset by finding existing data or scraping data.
2. Pre-processing of data (cleaning from unnecessary data, trim, regex, lowercase, tokenization, filtering of stop words).
3. Data polarization (word scoring).
4. Data aggregation (sorting, grouping, etc.).
5. Data presentation.

The authors used the TextBlob library to evaluate the data, as well as the Pearson correlation coefficient formula (see Equation 1), which compares two sets of data and measures the linear correlation between these data:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (1)$$

where r – Pearson's correlation coefficient; x and y – the values of two variables; $\bar{x}$ та $\bar{y}$ – the average values of x and y, respectively.

Unfortunately, the TextBlob library supports only English by default, so to support Ukrainian, we need to overwrite certain modules of the library.

VADER (Value Aware Dictionary and sEntiment Reasoner) is a tool for analyzing the sentiment of text developed for the Python programming language. It can be adapted to analyze emotional coloring, as it is based on a dictionary containing words with known emotional connotations [9].

In [10], the authors study customer reviews for specific domains based on the Improved VADER model. The proposed system for emotional text analysis is shown in Fig. 3.

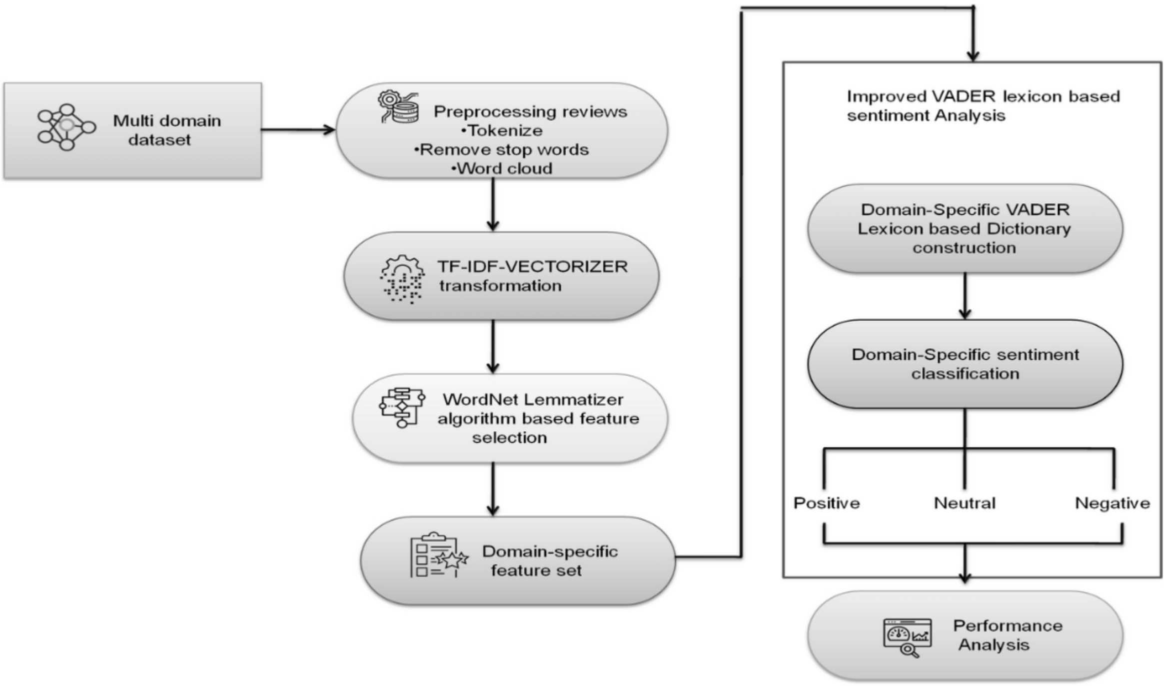

Fig. 3. Schematic of the text analysis system based on IVADER

Pre-processing is the standard procedure, but TF-IDF is much more interesting. This is an algorithm for determining whether a word belongs to a domain, using formulas that calculate the frequency of use of a word in a domain (see Equation (2)):

$$TF_{w_n} = \log \frac{g_{w_n}^{d_m}}{T_{d_m}}, \tag{2}$$

where TF_{w_n} – is the frequency of the term w_n in the document d_m, $g_{w_n}^{d_m}$ – the number of times the term w_n occurs in the document d_m, T_{d_m} – the total number of words in the document d_m.

And also the relation of the word to a specific domain (see Equation (3)):

$$IDF_{w_n} = \log(\frac{T_{d_m}}{N_{w_n}}), \tag{3}$$

where IDF_{w_n} – the weighted inverse document frequency for the term w_n; T_{d_m} – the total number of documents in the corpus in which the term w_n occurs; N_{w_n} – the number of documents in which the term w_n occurs.

The standard vaderSentiment library supports only English, but there are various options for translating text (using the APIs of various translators, such as DeepL or Google Translate) for use with vaderSentiment. However, such methods discard the peculiarities of natural language, which significantly reduces the accuracy of the result.

Word embeddings

Word embeddings are a word vectorization technique used in natural language processing (NLP). This technique allows representing words as vectors of numerical values that are used for further analysis of texts by machine algorithms. Word vectorization allows combining words by context and determine whether they belong to a particular phrase [11].

Word2Vec is a word embedding model developed by Google Research in 2013. This model is one of the first and most popular models for obtaining embedded word representations by vectors in low-dimensional space.

In [12], the authors try to analyze the contextual affiliation of words using graphs as an analog to classical vectors. Despite the fact that they used vector methods and skip-gram functions again, the accuracy of word evaluation is still slightly higher compared to classical text analysis tools. The graph was built using the NetworkX library. Word2Vec has certain disadvantages, namely:

1. It does not work well with morphologically rich languages (for example, in Ukrainian, these are different cases, genders, compound words). This can be solved with FastText.
2. Global information is not stored (using context from a set of sentences rather than just one sentence). This can be solved with GloVe.
3. Lack of a broad understanding of the context (for example, a lock can be both a fortress and a device that locks the door in Ukrainian. It is solved by transformers such as Bert, GPT).

FastText is a word embedding method developed by Facebook AI Research (FAIR) that extends the word2vec model by adding the ability to take into account the internal structure of a word (subword information). One of the main advantages of FastText is that it allows representing words as the sum of their subword vectors, which allows us to take into account morphological information and use embeddings even for unfamiliar words.

In this article [13], the authors explore the analysis of customer reviews for the Arabic language. The authors created 9 models based on deep learning, but used FastText for word embedding. FastText can be used to analyze the Ukrainian language, as the library officially provides a file for training an artificial intelligence model.

GloVe (Global Vectors for Word Representation) is a word embedding model developed by students at Stanford University in 2014. It is designed to obtain embedded word representations that reflect the semantic relationships between words in a text.

The authors of this article [14] analyze tweets and try to determine public attitudes toward ChatGPT. They use Word2Vec together with GloVe to find tweets that are similar in context. Somewhat similar to TF-IDF described above.

Supervised Learning

Machine learning methods, such as Supervised Learning, can be used to analyze texts emotionally. This approach means that the model is trained on labeled data, where each text has corresponding emotional labels (e.g., positive, negative, neutral). The models can use a variety of algorithms, such as logistic regression, support vector machine (SVM), neural networks, or decision trees, to classify texts according to their emotional connotations. Once trained, the models can be used to classify new texts according to their emotional characteristics [15].

Logistic Regression is a machine learning method used to solve classification problems. Despite its name, logistic regression is used to predict the probability that an input sample belongs to a certain class. This method is one of the simplest and most effective classification algorithms.

The authors of this article [16] analyze the news and, based on the emotional load of the message, predict changes in the rates of various cryptocurrencies. The data are vectorized and entered into a logistic regression model as training data (see Equation (4)):

$$P(y = 1) = \sigma(w \cdot x + b) = \frac{1}{1 + e^{-(w \cdot x + b)}}, \tag{4}$$

where $P(y = 1)$ – the probability that the output signal belongs to class 1; σ – a logistic function; w – a vector of weights for each input feature, x a vector of input features; b – a bias.

The results of other models are then compared.

The Support Vector Machine (SVM) method is a powerful machine learning algorithm used for both classification and regression. The main idea of the method is to find the optimal hyperplane that separates two classes of data so that it separates them from each other as much as possible.

In [17], the authors try to classify and categorize different types of text documents based on neural networks of different classifiers (machine learning methods).

As a result, the authors made a good comparison of different machine learning methods for text analysis (see Table 2).

Table 2

Comparison of machine learning methods with a teacher

Class/approach	Algorithms	Advantages	Disadvantages
Supervised learning	Support vector machine (SVM)	• SVM is capable of handling nonlinear decision boundaries. • Robust against over fitting issues. • Can work with large size data	• Large number of dimensions. • Difficulty in picking an efficient kernel function. • Time and memory complexity is high
	K-nearest neighbor (KNN)	• Effectiveness in text classification. • Non-parametric. • Handles multi-class data sets	• Computationally expensive. • Difficulties finding an optimal k value. • Challenging to find a meaningful distance function
	Decision tree (DT)	• Handles categorical features easily. • Divides hierarchically the data and works well with decision margins parallel to the feature axis. • Fast in learning and prediction	• Overfit. • Sensitive to perturbations in the data set. • The noise handling is bad
Ensemble learning	Random forest (RF)	• With decision tree ensembles, training time is reduced compared to other approaches. • There is less variance in trees. • The input data does not need to be prepared or pre-processed	• Slow predictions. • Large number of trees increases the difficulty of the prediction stage. • Visually, it is not as straightforward. • Overfitting is a common problem. • Choosing the right number of trees for a forest is necessary

Naive Bayes classifier is a simple but effective machine learning method, especially in text and review classification tasks. It is based on Bayes' theorem and is considered "naive" because it assumes independence between features (words) in the input data.

The authors of this article [18] used SVM and compared it with other classifiers, including Bayesian. For example, the graph of support vectors for the analyzed system (Fig. 4).

The authors used the formula of a naive Bayesian classifier (Form 5), which also tries to predict the future value based on trained data, but this classifier assumes data independence:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{5}$$

where $P(A|B)$ – the conditional probability of event A provided that event B occurs; $P(B|A)$ – the conditional probability of event B provided that event A occurs; $P(A)$ the a priori probability of event A; $P(B)$ – the a priori probability of event B.

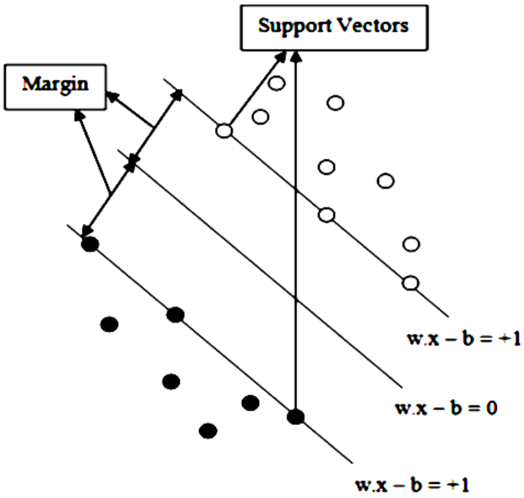

Fig. 4. Graph of SVM support vectors

A Decision Tree is a graphical structure that models decision making based on conditional branching. It is a tree-like structure where each node represents a test on a feature, each edge represents a test result, and each leaf corresponds to the final classified or regression result. Decision trees are widely used as simple and interpretable models for classification and prediction. In this article [19], the authors build a goal tree classifier for the system based on the formulas (Eq. (6)–(8)):

$$d_{ij} = \sqrt{\sum_{i=1}^{m} (x_i - x_j)}, \tag{6}$$

where d_{ij} – the distance between objects *i* and *j*, x_i and x_j – the feature values of objects *i* and *j*; *m* – the number of features.

$$d(t, s) = x_t + x_s, \tag{7}$$

where $d(t, s)$ – the distance between objects or points *t* and *s*; x_t and x_s – the values of the feature.

$$BTC(R) = -2\sum_{i=1}^{j} d_i + m_i \log(N), \tag{8}$$

where $BTC(R)$ – a lens that measures the cost or value of a function *R*.

As a result, the decision tree looks like the one shown in Fig. 5.

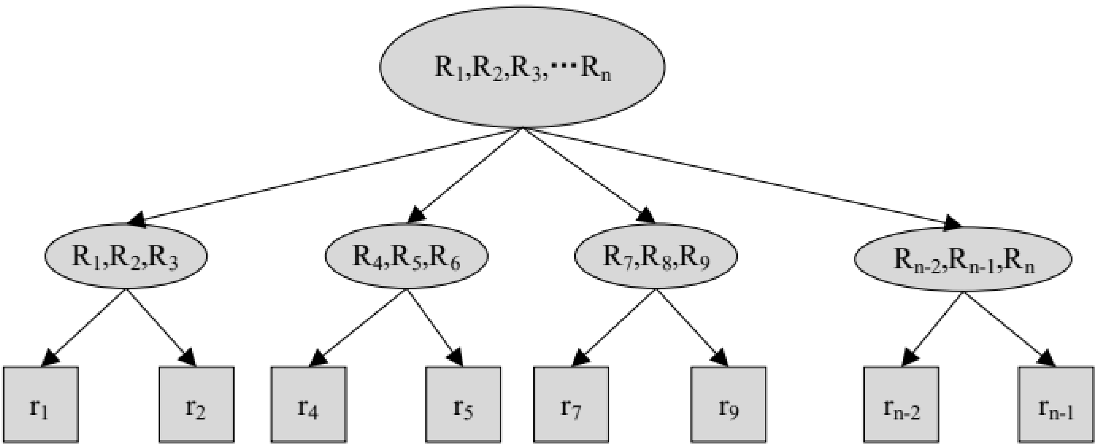

Fig. 5. The designed decision tree

Use of the analyzed methods for the Ukrainian language

After analyzing the existing machine learning methods for text analysis and emotional coloring, we can identify some of them that can be partially used to analyze Ukrainian-language text content. First of all, we can use the NRC EmoLex dictionary, as it officially supports Ukrainian localization. We can also use VADER algorithms, but not directly, since this tool does not support the Ukrainian language.

It is important to use word vectorization methods to find the context and relationship of key pairs. To do this, we can use word2vec, namely the FastText tool, which has an official support file for training models in Ukrainian. After the models have been formed and trained, it is necessary to proceed to the step of predicting the further result based on training. For this purpose, we can use various machine learning methods, but the most accurate ones are logistic regression and support vector machine.

The use of deep learning methods is unnecessary, as the complexity of the system does not require multi-level layers of neural networks.

As a result of the analysis of the methods, their classification was carried out, the result of which is presented in the form of a graph in Fig. 6.

This classification describes all possible options for emotional content analysis. After the analysis, certain categories that are not suitable for solving the task were identified and rejected, namely: The use of methods for analyzing non-textual content; use of deep multilevel learning; the use of methods that do not have native support for the Ukrainian language.

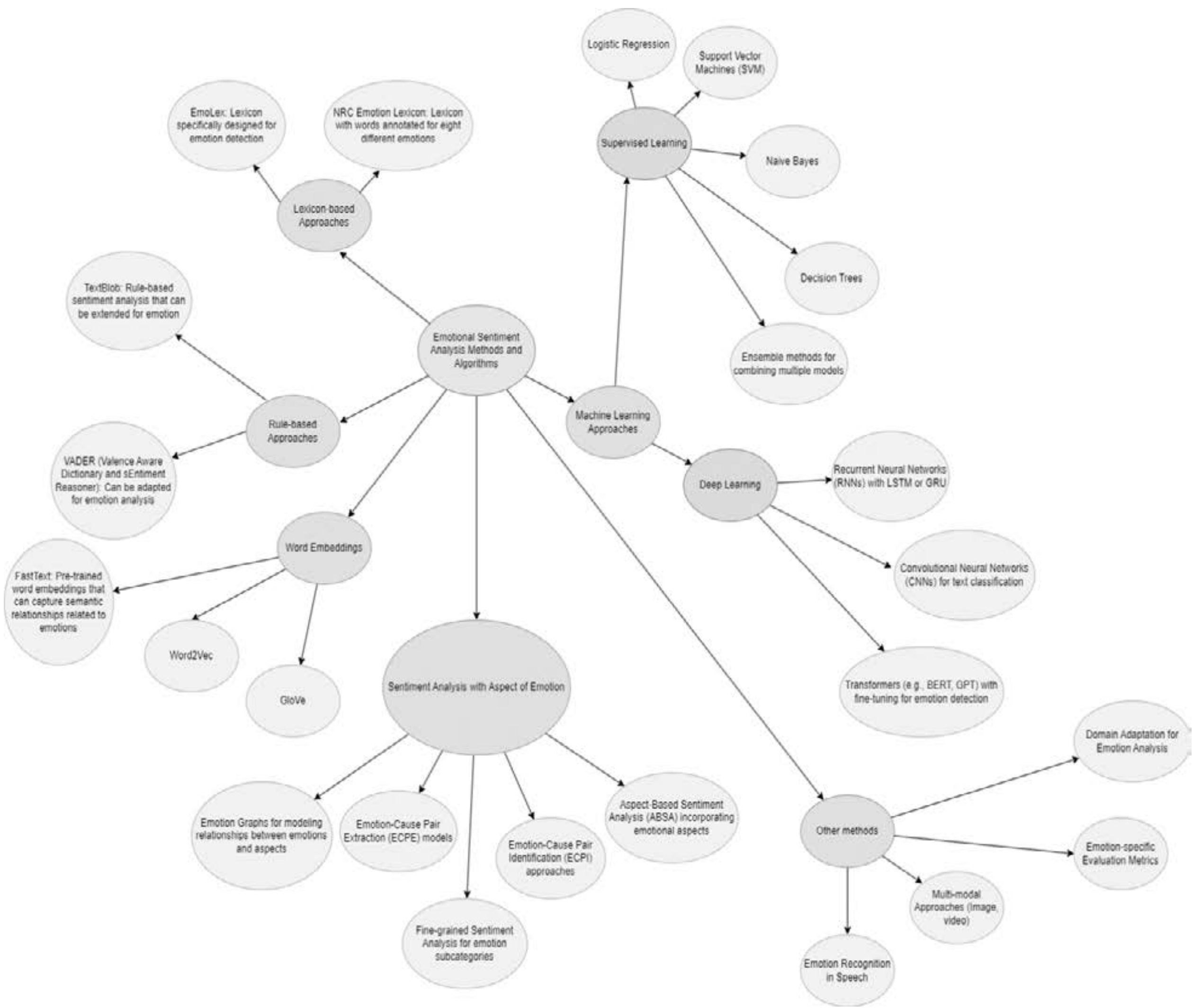

Fig. 6. Classification of methods for determining emotional coloration

The next step was to design the system using a structural approach. Structural design is an extremely important stage in the development of any system. Designing improves:

1. Organization and structure of the system parts.
2. Determining the level of scaling.
3. Readability and understanding of the system.
4. Efficiency and reduction of duplicates.
5. Testing and debugging.

Design of the system

The system design phase is a critical stage in the information systems development process. During this stage, developers thoroughly analyze user needs and business requirements, determine the structure of the system and its components, select technologies and tools for implementation, design user interfaces, and develop security and data protection strategies. Analyzing user needs is a primary task at the design stage. This analysis helps to understand how the system should work and what functions it should perform to meet user needs. Then, based on the requirements analysis, the system architecture is determined. An important part of this process is the selection of technologies and tools that will be used to implement the system, taking into account the needs and requirements of the project. During the system design, user interfaces are also developed. This is important to ensure convenient and efficient user interaction with the system. Interface development includes designing controls, organizing information flows, and displaying data to users. The IDEF0 standard was chosen for the system design.

For the designed system, a context diagram (see Fig. 7) was created that shows the main process of Analyze the emotional coloring of Ukrainian Language Content for Data Mining Systems.

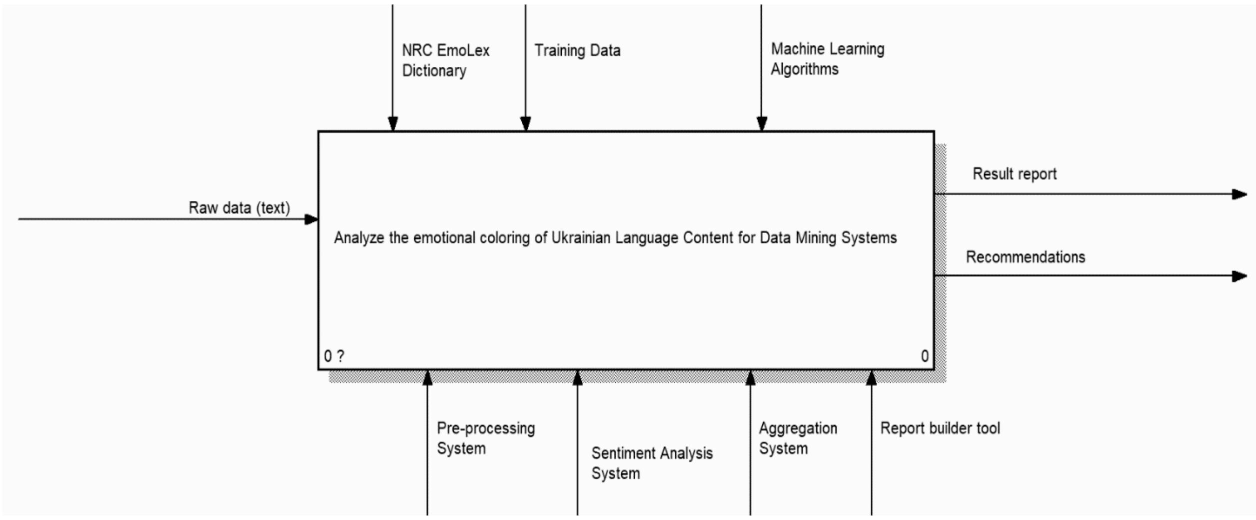

Fig. 7. Context diagram

The main process is – Analyze the emotional coloring of Ukrainian Language content. This process describes the main task of the system, which is performed through input, output, control flows, and mechanisms.

- Input data – Raw Data (text). This is content that needs to be analyzed.
- Output data – Result report, Recommendations. This is the result of the successful implementation of the main process.
- NRC EmoLex Dictionary, Training Data, Machine Learning Algorithms are used for management.
- Mechanisms – Pre-processing system, Sentiment Analysis System, Aggregation System, Report builder tool.

NRC EmoLex Dictionary is a dictionary described above. It allows you to evaluate each word by emotional spectrum and polarity.

Training Data – training data required to establish the correct operation of the neural network.

Machine Learning Algorithms – the algorithms described above for setting up network processes.

Pre-processing system – a system for pre-processing data, including trim, deduplicate, tokenize, etc.

Sentiment Analysis System – a system for analyzing the emotional spectrum based on the result of evaluating each word separately, as well as setting up the context.

Aggregation System – a system for grouping and sorting the resulting data.

In order to detail the main process, the context diagram was decomposed. The decomposed diagram is shown in Fig. 8.

The main processes are:

1. Pre-process data – processing of primary data (trim, stemming, tokenize).
2. Training of the sentiment analysis model (using the available dictionary and training data for the Ukrainian language).
3. Data analysis using classifiers (using machine learning algorithms for prediction based on training and real data).
4. Sorting and grouping the source data. Create a report (group data and analyze the result).

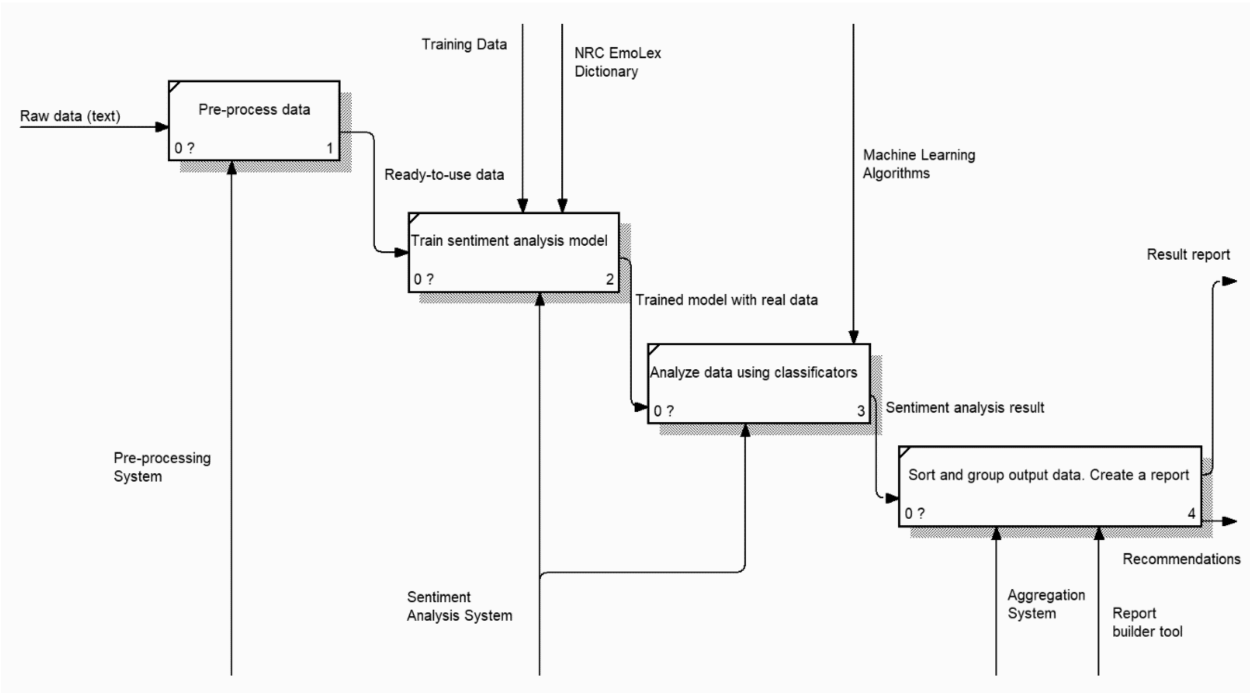

Fig. 8. Decomposed IDEF0 diagram

In general, the system design phase determines the fundamental aspects of its implementation, including its functionality, structure, and user experience.

Conclusion

The study analyzes the relevance of our subject area and shows its importance given the diversity of sources of Ukrainian-language content and the need for effective analysis of its emotional coloring. The article analyzes machine learning methods in detail, identifies their advantages and disadvantages, and the possibility of applying them to the model of emotional analysis of Ukrainian-language content. The result of this analysis is a structured classification of methods, which is displayed in the form of a graph. This classification can be used to analyze text and classify its emotional coloration. At the next stage of the study, the system was designed using the IDEF0 functional design methodology, which provided the necessary tools for further implementation of the system and its functionality.

Further research will be aimed at creating a prototype of the system for determining the emotional coloring of Ukrainian-language content and testing its operation. This prototype will be an important step in the improvement and development of data mining systems for Ukrainian-language content with emotional context.

References

1. Otamendi, J., F., & Martín, D. L. S. (2020, September 4). *The Emotional Effectiveness of Advertisement*. Frontiers. Retrieved February 28, 2024, from https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.02088/full

2. Basyuk, T., & Vasyliuk, A. (2023). Peculiarities of an Information System Development for Studying Ukrainian Language and Carrying out an Emotional and Content Analysis. *CEUR Workshop Proceedings, 3396*, 279–294.

3. Ou, L. C., Luo, M. R., Woodcock, A., & Wright, A. B. (2004). A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application, 29*(3), 232–240. https://doi.org/10.1002/col.20010

4. Bessa, A. (2023, December 11). *Lexicon-based sentiment analysis: What it is & how to conduct one*. Knime. Retrieved February 28, 2024, from https://www.knime.com/blog/lexicon-based-sentiment-analysis

5. Fagbola, T. M., & Surendra Colin, T. (2019). Lexicon-based bot-aware public emotion mining and sentiment analysis of the Nigerian 2019 presidential election on Twitter. *International Journal of Advanced Computer Science and Applications, 10*(10), 329–336. https://doi.org/10.14569/ijacsa.2019.0101047

6. Guzsvinecz, T., & Szűcs, J. (2023). Length and sentiment analysis of reviews about top-level video game genres on the steam platform. *Computers in Human Behavior, 149*, A107955. https://doi.org/10.1016/j.chb.2023.107955

7. Kirti, A. (2023, April 17). *Rule Based Approach in NLP*. *Geeksforgeeks*. Retrieved February 28, 2024, from https://www.geeksforgeeks.org/rule-based-approach-in-nlp/

8. Koukaras, P., Rousidis, D., & Tjortjis, C. (2023). Unraveling Microblog Sentiment Dynamics: A Twitter Public Attitudes Analysis towards COVID-19 Cases and Deaths. *Informatics, 10*(4), A88. https://doi.org/10.3390/informatics10040088

9. Pragnya, S. S. (2022, January 16). *VADER (Valence Aware Dictionary and sentiment Reasoner) Sentiment Analysis*. Medium. Retrieved February 28, 2024, from https://swayanshu.medium.com/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698

10. Barik, K., & Misra, S. (2024). Analysis of customer reviews with an improved VADER lexicon classifier. *Journal of Big Data, 11*(1), A10. https://doi.org/10.1186/s40537-023-00861-x

11. (n.d.). *A Guide on Word Embeddings in NLP*. Turing. Retrieved February 28, 2024, from https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

12. Moudhich, I., & Fennan, A. (2024). Graph embedding approach to analyze sentiments on cryptocurrency. *International Journal of Electrical and Computer Engineering, 14*(1), 690–697. https://doi.org/10.11591/ijece.v14i1.pp690-697

13. Hicham, N., Nassera, H., & Karim, S. (2024). Enhancing Arabic E-Commerce Review Sentiment Analysis Using a hybrid Deep Learning Model and FastText word embedding. *EAI Endorsed Transactions on Internet of Things, 10*. https://doi.org/10.4108/eetiot.4601

14. Su, Y., & Kabala, Z. J. (2023). Public Perception of ChatGPT and Transfer Learning for Tweets Sentiment Analysis Using Wolfram Mathematica. *Data, 8*(12), A180. https://doi.org/10.3390/data8120180

15. (n.d.). *What is Supervised Learning?* Google Cloud. Retrieved February 28, 2024, from https://cloud.google.com/discover/what-is-supervised-learning

16. Aysan, A. F., Caporin, M., & Cepni, O. (2024). Not all words are equal: Sentiment and jumps in the cryptocurrency market. *Journal of International Financial Markets, Institutions and Money, 91*, A101920. https://doi.org/10.1016/j.intfin.2023.101920

17. Labd, Z., Bahassine, S., & Housni, K. (2024). Ext classification supervised algorithms with term frequency inverse document frequency and global vectors for word representation: A comparative study. *International Journal of Electrical and Computer Engineering, 14*(1), 589–599. https://doi.org/10.11591/ijece.v14i1.pp589-599

18. Tabany, M., & Gueffal, M. (2024). Sentiment Analysis and Fake Amazon Reviews Classification Using SVM Supervised Machine Learning Model. *Journal of Advances in Information Technology, 15*(1), 49–58. https://doi.org/10.12720/jait.15.1.49-58

19. Liu, J., & Si, J. (2024). Digitization of Civics in College Physical Education Courses Based on the Correlation Matrix. *Applied Mathematics and Nonlinear Sciences, 9*(1). https://doi.org/10.2478/amns.2023.2.01576

МЕТОДИ МАШИННОГО НАВЧАННЯ ТА ПРОЄКТУВАННЯ СИСТЕМИ ВИЗНАЧЕННЯ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ УКРАЇНСЬКОМОВНОГО КОНТЕНТУ

Антон Ломовацький[1], Тарас Басюк[2]

Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж, Львів, Україна
[1] E-mail: anton.a.lomovatskyi@lpnu.ua, ORCID: 0009-0004-5170-3272
[2] E-mail: Taras.M.Basyuk@lpnu.ua, ORCID: 0000-0003-0813-0785

У статті автори проаналізували сучасний стан досліджень у галузі емоційного аналізу україномовного контенту для систем інтелектуального аналізу даних. Проаналізовано основні методи та підходи до вирішення проблеми. Також розглянуто основні алгоритми машинного навчання для аналізу текстового контенту. В результаті аналізу визначено основні методи і підходи, які можна використати для аналізу саме української мови, та здійснено їх класифікацію. Подальшим етапом стало проєктування функціональності системи з використанням структурного підходу. Розроблено контекстну діаграму інформаційної системи та здійснено декомпозицію її головного процесу, щоб детальніше відобразити процес підготовки та аналізу інформації під час визначення емоційного забарвлення.

Ключові слова: емоційний аналіз; машинне навчання; проєктування; класифікація методів; інформаційні технології.