DOI 10.31891/2307-5732-2024-343-6-22
УДК 004.415

LOMOVATSKYI ANTON
Lviv Polytechnic National University
https://orcid.org/0009-0004-5170-3272
e-mail: Anton.A.Lomovatskyi@lpnu.ua
BASYUK TARAS
Lviv Polytechnic National University
https://orcid.org/0000-0003-0813-0785
e-mail: Taras.M.Basyuk@lpnu.ua

NAIVE RULE-BASED METHOD IN SENTIMENT ANALYSIS OF UKRAINIAN-LANGUAGE CONTENT

This paper deals with the obstacles faced while executing the Naive Rule Based algorithms for analyzing sentiment of Ukrainian language content. Text sentiment analysis is useful to such types of content as feedbacks, brand tracking, political stances and psychological analyses. A number of steps, which include the previously described treatment of text, explication of elements of the text, emojis, links, hashtags and other special characters are abandoned within this work. The next step tackles 'tokenization' where the text is broken into a set of small units or words and 'lemmatization' where words are worn down to the basic word form for the purpose of analyses. After these steps are taken care of, a sentiment lexicon is used to classify the text in terms of its tone i.e. positive, negative and neutral. Even though it is very basic, the Naive Rule-Based method stands out for its simple and effective approach for carrying out sentiment analysis, ideal for scenarios in which more advanced and complex machine learning techniques may not be possible owing to data, computing power, and time limitations. This technique makes it possible to carry out easy customization and even use of domain specific language by simply expanding the sentiment lexicon or changing the rules. Nevertheless, there are some aspects in which the method falls short. More complex properties of language such as sarcasm, the meaning within context, and building up deep complex sentences are all aspects that the system can struggle with and therefore limit its accuracy. This study proposes some useful recommendations to address the limitations, including extending the existing emotion lexicon so more emotions can be comprehended, and the implementation of context-embedded methods. Further, hybrid methods that combine conventions and rules in addition to machine learning approaches are identified as prospects for improving the effectiveness. One such case study demonstrates the effectiveness of the Naive Rule-based approach as applied to the dataset in the Ukrainian language. The results demonstrate the capability of the method to provide clear sentiment scoring and emotion classification. Although this approach is not at the level of machine learning models, it still manages to be an efficient and feasible approach for specific use cases especially in cases where speed and clarity take precedence. The findings of this study emphasize that, in resource-poor settings, sentiment analysis can be carried out using this technique with data processing tools and methods that have low precision and context dimensions, and more ways to enhance accuracy and context dimensions are provided.

Keywords: sentiment analysis, Naive Rule-Based approach, Ukrainian language, text preprocessing, emotion detection, sentiment lexicon, natural language processing (NLP), data analysis.

ЛОМОВАЦЬКИЙ АНТОН, БАСЮК ТАРАС
Національний університет "Львівська політехніка"

НАЇВНИЙ МЕТОД ЗАСНОВАНИЙ НА ПРАВИЛАХ ПРИ АНАЛІЗІ НАСТРОЇВ УКРАЇНСЬКОМОВНОГО КОНТЕНТУ

У цій статті розглядаються труднощі, що виникають при застосуванні алгоритмів на основі наївних правил для аналізу настроїв українськомовного контенту. Аналіз настроїв є корисним для таких видів контенту, як відгуки, відстеження бренду, політичні позиції та психологічний аналіз. Описано етапи обробки тексту, такі як видалення зайвих елементів (емодзі, посилань, хештегів), токенізація і лематизація. Метод наївних правил виділяється простотою та ефективністю, особливо у випадках, коли застосування складніших моделей машинного навчання обмежене через ресурси. Водночас, цей метод має обмеження в обробці складних аспектів мови, таких як сарказм і контекст. Для підвищення точності пропонується розширення емоційного лексикону та використання гібридних методів, що поєднують правила з підходами машинного навчання. Тести на масивах даних показали, що цей підхід може бути ефективним для швидкого і зрозумілого аналізу настроїв, особливо в умовах обмежених ресурсів. Результати дослідження підкреслюють, що метод наївних правил, хоча і не досягає рівня моделей машинного навчання, є досить дієвим для швидкого аналізу текстів у певних умовах. Особливо це стосується ситуацій, коли ресурси для впровадження більш складних моделей обмежені, а також коли важливими є простота налаштування та швидкість виконання. Для подальшого розвитку цього підходу пропонується покращувати лексикон настроїв, включаючи більше емоційних станів, та впроваджування методів, що враховують контекст і складніші мовні структури. Використання гібридних рішень, які поєднують правила з машинним навчанням, відкриває нові можливості для підвищення точності аналізу, дозволяючи обробляти більш складні мовні конструкції та контексти, які не під силу суто правилозалежним методам.

Ключові слова: аналіз настроїв, наївний підхід на основі правил, українська мова, попередня обробка тексту, виявлення емоцій, лексика настроїв, обробка природної мови (NLP), аналіз даних.

Main research tasks and their significance

Analyzing the emotional which means of words, or emotional textual content evaluation, is of superb significance in numerous fields. Emotional evaluation lets in a higher information of the temper and feelings of individuals who engage with products, services, or content. This enables organizations and corporations to higher recognize the wishes and dreams of clients or users. By studying emotional reactions to an emblem or product, organizations can fast reply to poor remarks or criticism. This is critical for retaining emblem popularity and making sure a high-quality image. Knowing the emotional reactions of clients enables to become aware of hassle regions in

provider and paintings to enhance them [1].

Language is a provider of subculture and precise meanings. Analysis in Ukrainian permits to don't forget cultural and linguistic functions that have an effect on the translation of the text. This enables to keep away from misunderstandings and misinterpretation of the emotional coloring of words. For companies, corporations and the authorities' corporations working in Ukraine or with a Ukrainian audience, it's far essential to recognize the emotional reactions of Ukrainian consumers. This permits them to higher recognize the wishes and expectancies of the neighborhood population. The use of the Ukrainian language in numerous fields, together with era and research, contributes to its popularization and help as an essential device of communique and cultural identity. Summarizing, the main goal of the research is to conduct an emotional analysis of the text using known methods.

The main objectives of this study are focused on the evaluation of the Naive rule-based method for emotional analysis of the Ukrainian language. This task is important for several reasons:

1. Language features. The Ukrainian language has its own specific challenges in the process of emotional analysis due to its complex morphology, syntax, and limited availability of annotated linguistic resources. Solving these problems with a naive rule-based method helps to understand how such methods can be effectively applied in this language context.
2. Creating a lexicon of Ukrainian sentiment. The lexicon will serve as the basis for the rule-based method, allowing words and phrases to be categorized into positive, negative, and neutral sentiments. This lexicon is key because it provides the necessary linguistic data for the analysis.
3. Implementation of rule-based algorithms for determining the sentiment of a particular text. This task is important because it investigates the effectiveness of a rule-based system in sentiment analysis, especially in comparison to more complex machine learning approaches.
4. Evaluation and analysis of results. Evaluating the effectiveness of the proposed method is a key task. It involves understanding the potential and limitations of rule-based methods in the context of Ukrainian sentiment analysis.

Main research

Naive Rule-Based Approach in emotional text analysis is a method of analyzing texts to determine emotional content (or tone) based on simple rules and predefined vocabularies. This approach uses a set of rules and lexicons that determine the emotional coloring of words or phrases [2]. The main components are dictionaries in which each word is associated with a certain emotion or polarity (positive, negative, neutral). After applying the lexicons and rules, the algorithm calculates the overall emotional tone of the text [3]. Text analysis is divided into four stages: collecting data for analysis, data pre-processing, determining the polarity and emotional color of each word, summarizing and aggregating the results.

Collecting data for analysis

Data series is an essential step in carrying out emotional textual content evaluation the usage of the naive rule-primarily based totally absolutely method. First, the clean definition of the cause of the evaluation, for example, to apprehend purchaser feelings approximately a selected product. This allows to interest on applicable information sources, together with social media, purchaser reviews, news, or boards which may be applicable to the studies topic [4].

The following types of sources were selected for this example:

1. An array of tweets in Ukrainian with the words "добрий" or "поганий" in them. This is a large source of data with a naive pre-determined assessment of positive or negative.
2. Literature sources, namely:

2.1.The tragicomedy "One Hundred Thousand" by Ivan Karpenko-Kary - this work is naively pre-determined as generally positive.

2.2.The story "Ukraine on Fire" by Oleksandr Dovzhenko - this work is naively pre-determined as generally negative.

2.3.Bibliography of Mykola Khvylovy, namely 9 of his novels in the order of their publication for a general analysis of his work and changes in mood over time. These works have no preliminary assessment, so it is difficult to predict the results.

Data processing

Data preprocessing is an important step in the process of emotional text analysis, as it prepares raw text data for further analysis. This stage helps to improve data quality, reduce noise, and increase the accuracy of the final results. For example, the following sentence as a data preprocessing example is provided: "@користувач сьогодні чудовий день для поїздки в гори. Це втілить мої давні мрії! Ось посилання на маршрут: https://maps.com #щастя :)".

The first stage of data processing is to remove emojis, links, hashtags, and mentions of users in the text. This data does not carry any emotional coloring (except for emojis, but this analysis should also be approached with caution).

Mathematically, this operation can be represented as formula:

$$T' = T \setminus \{c \in T \mid UnicodeRange(c) \in [U + 1F600, U + 1F64F]\} \quad (1)$$

where:

T — input text,

T' — text after removing emojis,

c — character in text T,

UnicodeRange(c) — function that defines the range of Unicode for a character c.

After this stage, the sentence looks like this: “сьогодні чудовий день для поїздки в гори. Це втілить мої давні мрії! Ось посилання на маршрут: https://maps.com”.

Next, the links have to be removed from the text. They can be removed using RegExp (Regular Expression). This is a tool for searching and processing text by regular expressions. To remove links, this regular expression can be used: “http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F])”. That is, if an object with such a pattern is found in the input array, it will be filtered. Mathematically, this operation can be represented as formula:

$$T'' = T' \setminus R_{link} \quad (2)$$

where:

R_{link}— regular expression for links,

T'' — text after removing the links.

After this stage, the sentence looks like this: “сьогодні чудовий день для поїздки в гори. Це втілить мої давні мрії! Ось посилання на маршрут:”.

Hashtags and tags (the mentions of users) are removed in a similar way. Words that have # (for hashtags) or @ (for tags) at the beginning have to be filtered. The next stage of preprocessing is to remove special characters from the text. Any numbers and other special characters are filtered out of the input data set, as these characters will only interfere with the naive analysis. Mathematically, this operation can be represented as formula:

$$T''' = T'' \setminus \{c \in T'' \mid c\ starts\ with\ \#\ or\ @\} \quad (3)$$

where:

T''' — text after removing special symbols.

After this stage, the sentence looks like this: “сьогодні чудовий день для поїздки в гори Це втілить мої давні мрії Ось посилання на маршрут”.

Next, the words need to be segmented, i.e., sentences and individual words need to be separated from each other. For a naive analysis, simply filtering out any punctuation marks is enough, since the context is not important in this case. This is a more programmatic transformation of the input data from a single string to an array of words for further processing.

After this stage, the sentence looks like this: “[['сьогодні', 'чудовий', 'день', 'для', 'поїздки', 'в', 'гори', 'Це', 'втілить', 'мої', 'давні', 'мрії', 'Ось', 'посилання', 'на', 'маршрут']]”.

This process is also called tokenization because each word is converted into a separate token, an element for analysis. The next step is lemmatization of the text [5]. This is the process of text normalization, which consists in reducing words to their basic or dictionary form, called a lemma. Mathematically, the lemmatization operation can be represented as formula:

$$L(w) = lemma(w) \quad (4)$$

where:

L(w)— lemma for word w,

lemma(w)— function that defines lemma for w.

Thus, the lemmatization of the text can be represented as formula:

$$Tlemma = \{L(w) \mid w \in T'''\} \quad (5)$$

After this stage, the tokens can be reduced to the form of a sentence. After this stage, the sentence looks like this: “сьогодні чудовий день для поїздка в гора це втілити мій давній мрія ось посилання на маршрут”. The next step is to filter out stop words in the sentence. Stop words are words that do not carry sentiment, that is, they are neutral and unnecessary when analyzing a sentence.

A stop word dictionary is used for filtering. A stop word dictionary is a predefined list of words that do not carry an important semantic load.

After this stage, the sentence looks like this: “чудовий день поїздка гора втілити давній мрія посилання маршрут” (The words “сьогодні, для, в, це, мій, ось” are stop words and do not carry any meaning).

Determination of polarity and emotional coloring

After data processing, the input was transformed from “@користувач сьогодні чудовий день для поїздки в гори. Це втілить мої давні мрії! Ось посилання на маршрут: https://maps.com #щастя :)” to “чудовий день поїздка гора втілити давній мрія посилання маршрут”. This data can be analyzed and emotional analysis can be performed.

Polarity analysis is an approach in the field of natural language processing (NLP) and sentiment analysis that aims to determine the emotional tone of a text. It involves determining whether a given text is positive, negative, or neutral [6, 7]. This algorithm also uses a dictionary of pre-rated words with scores of -1, 0, 1, respectively, for negative, neutral, and positive. After analyzing each word in the input sentence, we can come to the following result (Table 1):

Table 1

Polarity analysis of an example input data

Word	Estimation
чудовий	1
день	0
поїздка	0
гора	0

Word	Estimation
втілити	1
давній	0
мрія	1
посилання	0
маршрут	0

Next, summarization of the scores of each word is required and the result is 3. If the number is greater than zero, the sentence is positive. The degree of positivity depending on how much the score is greater than zero can be determined. Mathematically, the polarity analysis can be represented as the formula:

$$P = \sum_{i=1}^{n} S(w_i) \quad (6)$$

where:

- P— text polarity,
- $S(w_i)$— estimation for word w_i,
- n — the amount of words in the text.

The next type of analysis is the analysis of emotional coloring using the EmoLex dictionary [8]. After analyzing each word in the input sentence, the following result is presented in table 2. Summing up the scores for each emotion, we can see that the sentence is more of a surprise (2 points), as well as joyful, anticipatory and trusting. “чудовий” also has a negative connotation, but it does not affect the overall analysis of the sentence. Mathematically, the analysis of emotional coloring can be represented as formula:

$$E_k = \sum_{i=1}^{n} e_k(w_i) \quad (7)$$

where:

- E_k— estimation of emotion k,
- $e_k(w_i)$— estimation of emotion of the word w_i,
- n — the amount of characters in text.

Table 2

EmoLex analysis of an example input data

Word	anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	trust
чудовий	0	0	0	0	1	0	1	0	1	1
день	0	0	0	0	0	0	0	0	0	0
поїздка	0	0	0	0	0	0	0	0	1	0
гора	0	1	0	0	0	0	0	0	0	0
втілити	0	0	0	0	0	0	0	0	0	0
давній	0	0	0	0	0	1	0	0	0	0
мрія	0	0	0	0	0	0	0	0	0	0
посилання	0	0	0	0	0	0	0	0	0	0
маршрут	0	0	0	0	0	0	0	0	0	0

In general, the algorithm of the method can be depicted using a Petri net (Fig. 1).

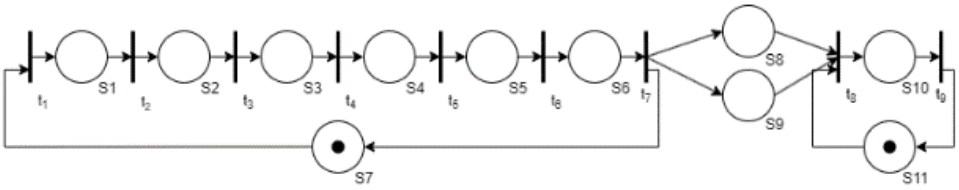

Fig. 1. Petri net that illustrates the work of the method

The purpose of each position and transitions is shown in Tables 3 and 4.

Table 3

Positions of Petri net

Position	Purpose
S1	Delete emoji
S2	Delete links
S3	Delete hashtags and tags
S4	Remove special characters
S5	Tokenization
S6	Lemmatization
S7	Pre-processing of all incoming data by the system

Position	Purpose
S8	Determination of polarity
S9	Determination of emotional coloring
S10	Data aggregation
S11	Creation of graphs by the system

Table 4

Transitions of Petri net

Position	Purpose
t_1	Input text
t_2	Text after removing emoji
t_3	Text after removing links
t_4	Text after removing hashtags and tags
t_5	Text after removing special characters
t_6	Array of tokens
t_7	Array of lemmas
t_8	Processed data with scores
t_9	The result in the form of graphs

Analyzing a data set using a naive rule-based approach

Now let's move on to the analysis of the literary sources specified in the collecting data stage. For the first test, tweets were selected by filtering for the presence of the emotion-containing words. These datasets were further processed, evaluated, and aggregated. For the dataset containing the word "добре", 6565 tweets were analyzed and emotional coloring was determined by polarity (Fig. 2) and using EmoLex (Fig. 3).

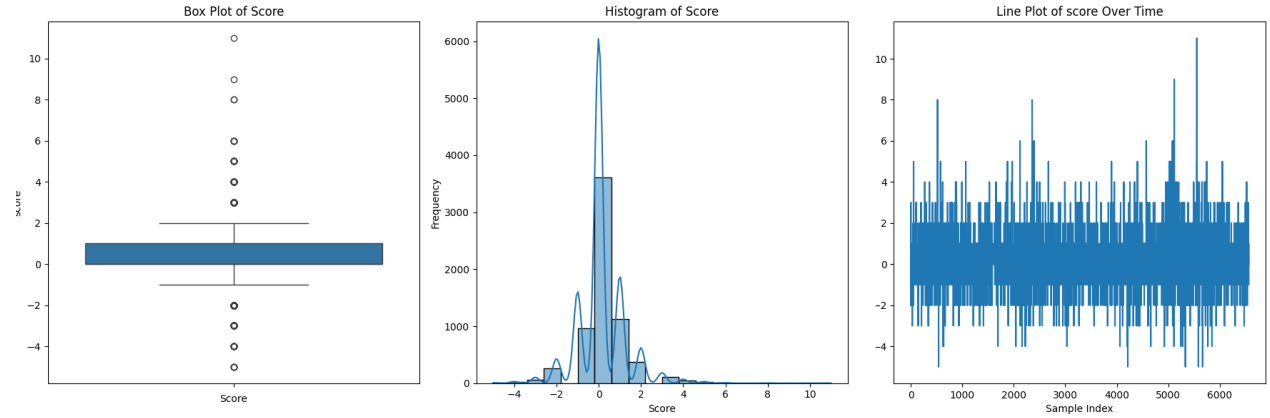

Fig. 2. Polarity analysis of the dataset "добре"

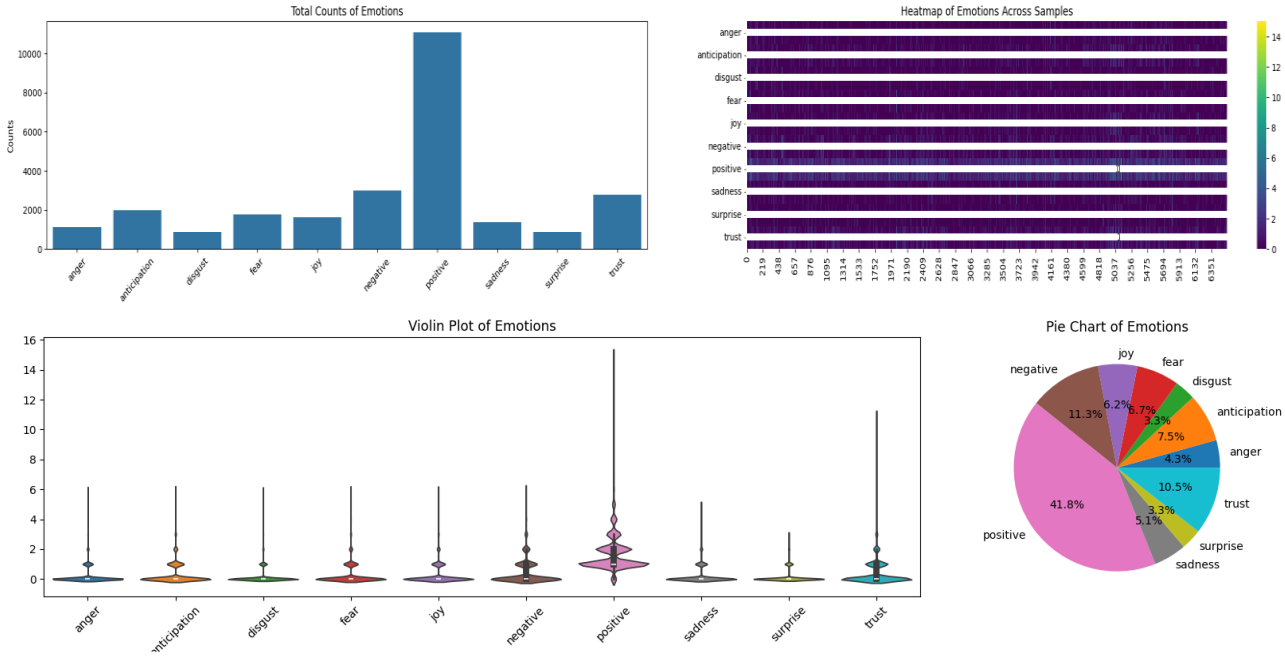

Fig. 3. EmoLex analysis of the dataset "добре"

For the “добре” dataset, the positive emotion prevails by a large margin, and the polarity is on the verge of 0-1, so, in general, the result of the analysis can be considered successful.

Similar experiments to determine emotional coloration by polarity and using EmoLex were performed for the described datasets:

- For the dataset containing the words “добре” and “погано”, 6953 tweets were analyzed, with the positive emotion prevailing by a wide margin, but other negative emotions are also present, and the polarity is at the 0 limit.
- For the dataset containing the words “добрий”, “погано”, “поганий”, 7294 tweets were analyzed, with positive and negative emotions being almost at the same level, and polarity being on the borderline of -1 and 1.
- For the dataset containing the word “погано”, 1917 tweets were analyzed, with the emotion positive prevailing by a large margin, but other negative emotions such as fear, expectation, and negativity exceed the positive rating in total, and are on the verge of 0-1 in terms of polarity.
- For the tragicomedy “One Hundred Thousand” by Ivan Karpenko-Kary a significant range of emotions has been identified, where positive emotions still prevail, but the work is defined as negative in terms.
- For the novel “Ukraine on Fire” by Oleksandr Dovzhenko a significant range of emotions has been identified, where negative emotions still prevail, but the polarity of the work is sharply negative.
- For the bibliography of 9 stories by Mykola Khvylovy, a significant range of emotions is identified, where in general they are on the same level. In terms of polarity, all the works are defined as negative, but the tendency of the author's mood changes is interestingly depicted.

The result of the allocation of resources and time to perform calculations is shown in Fig. 4:

```
Execution time: 2992.2549204826355 seconds
CPU time (user): 2942.984375 seconds
CPU time (system): 36.3125 seconds
Memory usage: 3763.1015625 MB
```

Fig. 4. The result of the allocation of resources and time

Conclusion

This article analyzes the importance of analyzing textual Ukrainian-language content using the example of the naive rule-based method of emotional analysis. Based on the study, various emotional trends were identified in different datasets. In most cases, the analysis demonstrates the prevalence of certain emotions, such as positive or negative, depending on the context. However, it is worth noting that the results are not always unambiguous: in some cases, the polarity of the emotional assessment is near zero (the “good and bad” dataset), which indicates a balance between positive and negative emotions, while in other cases, emotional assessments lean more towards one side. In general, the analysis revealed important emotional patterns, although some results require further refinement to more accurately identify emotional characteristics in different texts. After the analysis, it can be concluded that the method works, but with inaccuracies. Areas for improving the emotional coloration algorithm can be identified, such as: correctly selected input data and their volume, moving away from the naivety of the method, moving away from rules and dictionaries.

Further research will be aimed at improving the model using artificial intelligence methods.

References

1. R. Strubytskyi and N. Shakhovska, "Method and models for sentiment analysis and hidden propaganda finding," Computers in Human Behavior Reports, vol. 12, art. 100328, Dec. 2023, doi: 10.1016/j.chbr.2023.100328.

2. O. Mediakov and T.Basyuk, “Specifics of Designing and Construction of the System for Deep Neural Networks Generation”, CEUR Workshop Proceedings, Vol-3171, 2022, pp.1282–1296.

3. G.K. Wadhwani, P.K. Varshney, A. Gupta, and S. Kumar, "Sentiment Analysis and Comprehensive Evaluation of Supervised Machine Learning Models Using Twitter Data on Russia–Ukraine War," SN Computer Science, vol. 4, no. 4, art. 346, Jul. 2023, doi: 10.1007/s42979-023-01790-5.

4. Guerra, "Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict," Frontiers in Artificial Intelligence, vol. 6, art. 1163577, 2023, doi: 10.3389/frai.2023.1163577.

5. L. Y. Madhavi and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," ETRI Journal, vol. 46, no. 4, pp. 633-647, Aug. 2024, doi: 10.4218/etrij.2023-0162.

6. M. Alfreihat, O. S. Almousa "Emo-SL Framework: Emoji Sentiment Lexicon Using Text-Based Features and Machine Learning for Sentiment Analysis," IEEE Access, vol. 12, pp. 81793-81812, 2024, doi: 10.1109/ACCESS.2024.3382836.

7. C. Pipal, B. N. Bakker, G. Schumacher, and M. A. C. G. van der Velden, "Tone in politics is not systematically related to macro trends, ideology, or experience," *Scientific Reports*, vol. 14, no. 1, Art. no. 3241, Dec. 2024, doi: 10.1038/s41598-023-49618-9.

8. T. M. Fagbola, "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, vol. 10, no. 10, pp. 329-336, 2019, doi: 10.14569/ijacsa.2019.0101047.