UDC 004.415
DOI DOI DOI https://doi.org/10.32782/2663-5941/2026.2.1/25

Lomovatskyi A.A.
https://orcid.org/0009-0004-5170-3272
Lviv Polytechnic National University

Basyuk T.M.
https://orcid.org/0000-0003-0813-0785
Lviv Polytechnic National University

METHODS FOR SENTIMENT ANALYSIS OF UKRAINIAN-LANGUAGE CONTENT: RULE-BASED, MACHINE LEARNING AND TRANSFORMER APPROACHES

Sentiment analysis is integral to contemporary natural language processing, as it autonomously identifies emotional tone within textual data across diverse domains such as social media, decision-making, marketing, and public opinion. Nonetheless, conducting sentiment analysis in morphologically rich, low-resource languages like Ukrainian presents significant challenges due to limited annotated datasets, linguistic diversity, flexible syntactic structures, and complex grammatical frameworks. These obstacles hinder the direct application of methods originally devised for English and other high-resource languages.

This article examines recent sentiment analysis techniques for Ukrainian content, emphasising rule-based systems, traditional machine learning, and transformer-based language models. It outlines their fundamental principles, highlighting their benefits, challenges, and suitability for Ukrainian texts, particularly in resource-scarce environments. The discussion considers rule-based methods as a practical starting point, machine learning models as adaptable classifiers, and transformer models as sophisticated tools capable of capturing contextual nuances and complex semantics.

This paper provides a brief comparison of the performance of various methods on Ukrainian-language data, employing standard metrics such as accuracy, precision, recall, and F1-score. The results demonstrate that, although rule-based methods continue to be valuable in specific contexts and serve as benchmarks, transformer models attain higher accuracy, exhibit improved generalisation, and show greater robustness to linguistic diversity. These findings assist in shaping methodologies for sentiment analysis in Ukrainian content and highlight potential opportunities for hybrid techniques to enhance low-resource language processing.

Keywords: *sentiment analysis, Ukrainian language, natural language processing, rule-based methods, machine learning, transformer models, low-resource languages.*

Formulation of the problem. Amid the rapid growth of textual data, sentiment analysis has become a vital part of modern natural language processing, enabling the automated assessment of emotional tone in user-generated content across social media platforms and within analytical systems. The ability to automatically recognise sentiment polarity allows organisations to monitor public opinion, analyse large-scale textual data, and enhance decision-making processes across various fields.

Although there has been notable progress in sentiment analysis for high-resource languages, applying modern methods to Ukrainian-language content remains challenging. These issues stem from the limited availability of annotated datasets, the morphological complexity of the Ukrainian language, flexible word order, and the variability of syntactic structures [1]. Consequently, many approaches designed for English and other high-resource languages cannot be directly adapted to Ukrainian without making methodological adjustments.

Modern sentiment analysis methods are usually categorised into rule-based systems, traditional machine learning techniques, and transformer-based language models. Each of these methods offers its own benefits and drawbacks when used with Ukrainian texts, especially in low-resource settings where annotated data and computational power are limited.

Therefore, an important scientific problem is to systematise contemporary approaches to sentiment analysis of Ukrainian-language content, identify their advantages and limitations, and assess their applicability in low-resource conditions. Addressing this problem contributes to the development of effective intelligent systems for Ukrainian-language text processing and supports further advancement of natural language processing methods for morphologically rich languages [2].

Analysis of recent research and publications. Recent advancements in sentiment analysis reveal considerable progress across various methodological paradigms, including rule-based approaches, traditional machine learning techniques, and transformer-based language models [3]. Early sentiment analysis systems mainly relied on lexicon- and rule-based methods, which remain pertinent for their interpretability and stability in low-resource environments. These approaches typically utilise sentiment dictionaries, polarity scoring mechanisms, and linguistic rules to determine sentiment orientation, making them especially appropriate for morphologically rich languages where annotated data are limited [4].

Recent progress in statistical natural language processing has led to the widespread use of machine learning techniques like Naïve Bayes, Support Vector Machines, and logistic regression for sentiment analysis. These approaches facilitate automated feature extraction and scalable classification but depend heavily on large, annotated datasets, which remain limited for Ukrainian-language content [5]. Consequently, their effectiveness can vary significantly depending on the domain and dataset size.

Recently, transformer-based language models like BERT and its variants have achieved leading results in sentiment analysis by effectively capturing context and semantic links in the text. Multilingual transformer models also support cross-lingual transfer learning, which benefits low-resource languages. However, deploying these models typically demands significant computational resources and domain-specific fine-tuning to perform well in Ukrainian-language environments [6].

Although there has been an increase in sentiment analysis studies, research remains fragmented in methodology, especially for Ukrainian-language content. Most works focus on evaluating individual approaches rather than providing a broad comparison across methods. Additionally, challenges such as linguistic variability, limited resources, and domain adaptation are not sufficiently explored in the existing literature [7].

Thus, additional research is needed to organise current sentiment analysis techniques and evaluate their effectiveness for Ukrainian-language texts, particularly in low-resource environments.

Task statement. The purpose of this study is to organise contemporary approaches to sentiment analysis and assess their suitability for Ukrainian-language content in low-resource conditions. The study aims to identify methodological strengths and weaknesses of different approaches and determine their effectiveness in processing morphologically rich texts.

To achieve this goal, the following objectives are defined:

– to analyse existing methodological approaches to sentiment analysis, including rule-based, machine learning, and transformer-based methods.

– to generalise the theoretical foundations of sentiment analysis for Ukrainian-language text processing.

– to organise the advantages and limitations of modern sentiment analysis approaches in low-resource environments.

– to identify key challenges related to applying modern sentiment analysis techniques to morphologically rich languages.

– to justify the feasibility of combining methodological approaches for enhancing sentiment classification performance.

Outline of the main content of the study. This research formalises sentiment analysis of Ukrainian-language texts as a supervised multi-class classification problem within a unified decision-making framework. The set of textual inputs is represented as a finite collection of natural-language messages, where each instance corresponds to a single text, and the overall dataset consists of N observations.

$$X = \{x_1, x_2, \ldots, x_N\} \tag{1}$$

where x_i – individual text, N – total number of texts.

Each text can be represented as an ordered sequence of linguistic units (2), reflecting its internal syntactic and lexical structure [8].

$$x_i = \left(w_1, w_2, \ldots, w_{L_i}\right) \tag{2}$$

where w_j – word/token, L_i – length of the text.

The sentiment classification task is defined over a finite label space (3):

$$Y = \{y_1, y_2, \ldots, y_{K\}} \tag{3}$$

where K is the number of sentiment classes.

This study utilises a three-class setup (4), representing positive, negative, and neutral sentiments. This method frames sentiment analysis as a categorical decision-making process, well-suited to applied information systems that require clear, interpretable results.

$$Y = \left\{text_{negative}, text_{positive}, text_{neutral}\right\} \tag{4}$$

The annotated dataset utilised for modelling can be represented as a set of input–output pairs (5), where each textual instance is linked with a predefined sentiment label [9].

$$D = \left(x_i, y_i\right)_{i=1}^{N} \quad (5)$$

where $x_i \in X$ – text instance, $y_i \in Y$ – corresponding sentiment label.

Within this formulation, sentiment analysis can be conceptualised as a functional mapping from the domain of textual inputs to the domain of sentiment labels (6).

$$f : X \to Y \quad (6)$$

where f is the sentiment classification function.

For data-driven approaches, this mapping is parameterised (7), allowing models to learn decision boundaries from empirical observations.

$$f_{\{\theta\}(x)} = \hat{y} \quad (7)$$

where θ – model parameters, $\hat{y}$ – predicted sentiment label.

From a modelling perspective, sentiment classification typically comprises two stages: representation learning and decision-making. The representation stage converts textual inputs into structured formats suitable for automated processing, while the classification stage assigns sentiment labels based on learned or predefined decision rules (8).

$$f_{\{\theta\}(x)} = g_{\theta(\varphi(x))} \quad (8)$$

where g_θ – classifier operating on representations.

The final decision is made using a maximum a posteriori criterion (9), ensuring consistent label selection across modelling paradigms.

$$\hat{y} = arg\,max_{y \in Y p(y \mid x, \theta)} \quad (9)$$

where $p(y \mid x, \theta)$ – predicted probability of class y, $\hat{y}$ – final predicted sentiment label.

This formalisation provides a unified conceptual basis for evaluating various sentiment analysis methods within consistent methodological conditions. Using a shared decision framework, differences in empirical outcomes are primarily attributed to modelling features rather than inconsistencies in task definition.

To facilitate an objective and consistent comparison across various sentiment analysis techniques, this study uses a standardised evaluation framework that standardises all phases of the experimental process. This framework establishes uniform methodological conditions for rule-based, traditional machine learning, and transformer-based models, allowing differences in performance to be mainly ascribed to the models' features rather than to variations in experimental setup [10].

The framework's overall structure is shown in Fig. 1. It comprises five interconnected layers: unifying datasets and preprocessing; adapting representations across different modelling paradigms; a unified protocol for training and fine-tuning; a standardised evaluation process; and a comparative analysis with reporting. This layered design ensures methodological consistency and encourages reproducible experiments across all approaches.

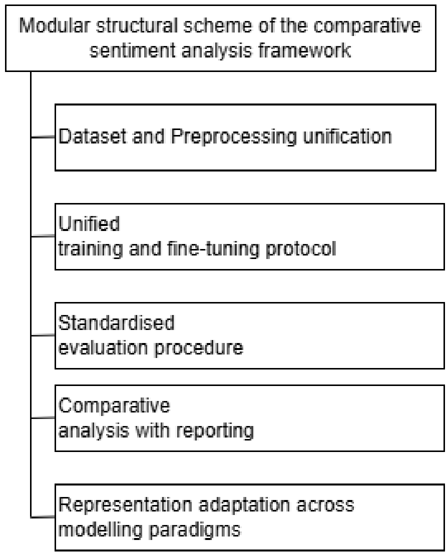

Fig. 1. Modular structural scheme of the comparative sentiment analysis framework

Within the proposed framework, all models are evaluated using a consistent annotated dataset, standardised sentiment label definitions, and a uniform data partitioning strategy. The dataset is divided into mutually exclusive training and testing subsets (10–11), with their proportions set by a fixed split coefficient (12). This method prevents information leakage and promotes fairness across different modelling approaches, particularly for data-driven methods trained on labelled data.

$$D = D_{train} \cup D_{test} \quad (10)$$

where D_{train} – training subset, D_{test} – testing subset.

$$D_{train} \cap D_{test} = \varnothing \quad (11)$$

where $\varnothing$ – denotes the empty set.

$$\left|D_{train}\right| = \alpha N,$$

$$\left|D_{test}\right| = \left(1 - \alpha\right) N \quad (12)$$

where $\alpha \in (0,1)$ – dataset split coefficient, N – total number of samples.

To ensure comparability, a standardised set of evaluation metrics is used across all methods. Predicted sentiment labels serve as classification outputs

(13), with ground-truth labels defined similarly (14). The classification results are presented as label pairs (15), enabling systematic differentiation between correct and incorrect predictions via an indicator-based approach (16). The overall predictive accuracy is then measured with a unified statistical metric (17).

$$\hat{Y} = (\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_M}) \tag{13}$$

where $\widehat{y_i}$ – predicted sentiment label, M – number of test samples.

$$Y = (y_1, y_2, \dots, y_M) \tag{14}$$

where y_i – true sentiment label for sample i.

$$(y_i, \widehat{y_i}) \tag{15}$$

where y_i – true label, $\widehat{y_i}$ – predicted label.

$$\mathbb{I}(y_i = \widehat{y_i}) \tag{16}$$

where $\mathbb{I}(\cdot) = 1$ if the condition holds, $\mathbb{I}(\cdot) = 0$ otherwise.

$$\mathbb{E}[\mathbb{I}(y = \hat{y})] \tag{17}$$

where $\mathbb{E}[\cdot]$ – expectation operator.

By maintaining consistent experimental conditions, this framework provides a reliable foundation for comparing sentiment analysis methods. It ensures reproducible results, facilitates clear interpretation of performance, and enables systematic assessment of sentiment analysis techniques for Ukrainian-language systems in low-resource settings.

The empirical evaluation is conducted on a Ukrainian sentiment corpus containing 9,161 short texts from social media. Positive and negative samples are naturally occurring user posts, whereas the neutral class includes brief factual statements that differ structurally from emotionally expressive content. The sentiment category distribution is displayed in Table 1.

Table 1

Full corpus distribution

Class	Train (80%)	Test (20%)	Total	Share
Positive	5,251	1,314	6,565	71.7%
Negative	1,677	419	2,096	22.9%
Neutral	400	100	500	5.5%
Total	7,328	1,833	9,161	100%

To maintain methodological consistency, the corpus is split into training and testing subsets through stratified sampling with a fixed random seed. This approach preserves class proportions and ensures fair comparison across different modelling paradigms. The annotated dataset facilitates evaluation of sentiment classification under consistent experimental conditions within a unified comparative framework.

Before modelling, texts undergo standard preprocessing steps. These include normalising whitespace, removing extra spaces, and replacing typographic symbols with ASCII equivalents. No lowercasing, stop-word filtering, or lemmatisation is applied, as these processes can distort lexical signals crucial for rule-based and TF-IDF methods and have limited advantages for subword transformer tokenisers.

All models are trained and evaluated under consistent experimental conditions. Classical machine learning baselines are trained on the full training subset, whereas transformer-based models are fine-tuned with a fixed hyperparameter set. A standardised training protocol is employed to facilitate cross-architecture comparisons, using uniform optimisation settings, batch sizes, sequence length limits, and reproducibility measures.

This comparative evaluation examines the performance of rule-based, classical machine-learning, and transformer-based sentiment analysis techniques within a single experimental framework. It aims to identify overall performance trends, analyse class-specific behaviour, and assess the relative effectiveness of various modelling approaches for sentiment classification in Ukrainian.

Model performance is evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score, all calculated under consistent conditions across approaches. The confusion matrix, as defined in (18), helps analyse the distribution of correct and incorrect predictions among sentiment classes. Additional metrics, such as true positives, false positives, and false negatives, calculated according to (19-21), enable detailed evaluation at the class level.

$$C = [c_{ij}] \tag{18}$$

where c_{ij} – number of samples whose true sentiment class is i and which are predicted as class j.

$$TP_k = c_{kk} \tag{19}$$

where TP_k – true positives for sentiment class k.

$$FP_k = \sum_{i \neq k} c_{ik} \tag{20}$$

where FP_k – number of samples incorrectly predicted as class k.

$$FN_k = \sum_{j \neq k} c_{kj} \tag{21}$$

where FN_k – number of samples of class k incorrectly predicted as other classes.

The overall classification results reveal consistent performance differences across different modelling paradigms. Transformer-based models perform best, given their ability to capture contextual dependencies, implicit sentiment cues, and intricate semantic relationships. Classical machine learning techniques also perform well in controlled environments, especially when lexical signals closely match sentiment polarity. Conversely, rule-based methods tend to lag, as they rely heavily on fixed lexical resources and have limited ability to model context.

Table 2 summarises the quantitative comparison of the evaluated models. The results indicate that data-driven approaches clearly outperform purely rule-based systems across all evaluation metrics. These performance differences are particularly evident in macro-averaged measures, which provide a more balanced view of classification across sentiment categories.

Table 2

Comparison of sentiment analysis models

Model	Accuracy	Precision (M)	Recall (M)	F1 (M)	Precision (m)	Recall (m)	F1 (m)
3_tfidf_logreg	0.98036	0.975485	0.98815	0.981499	0.98036	0.9803	0.98036
2_custom_ua_hybrid	0.971631	0.950001	0.98247	0.965465	0.971631	0.9716	0.971631
4_Geotrend_ distilbert-base-uk-cased	0.962357	0.9516	0.97924	0.964041	0.962357	0.9623	0.962357
1_ua_lexicon_ rulebased	0.749591	0.611506	0.69611	0.61494	0.749591	0.7495	0.749591

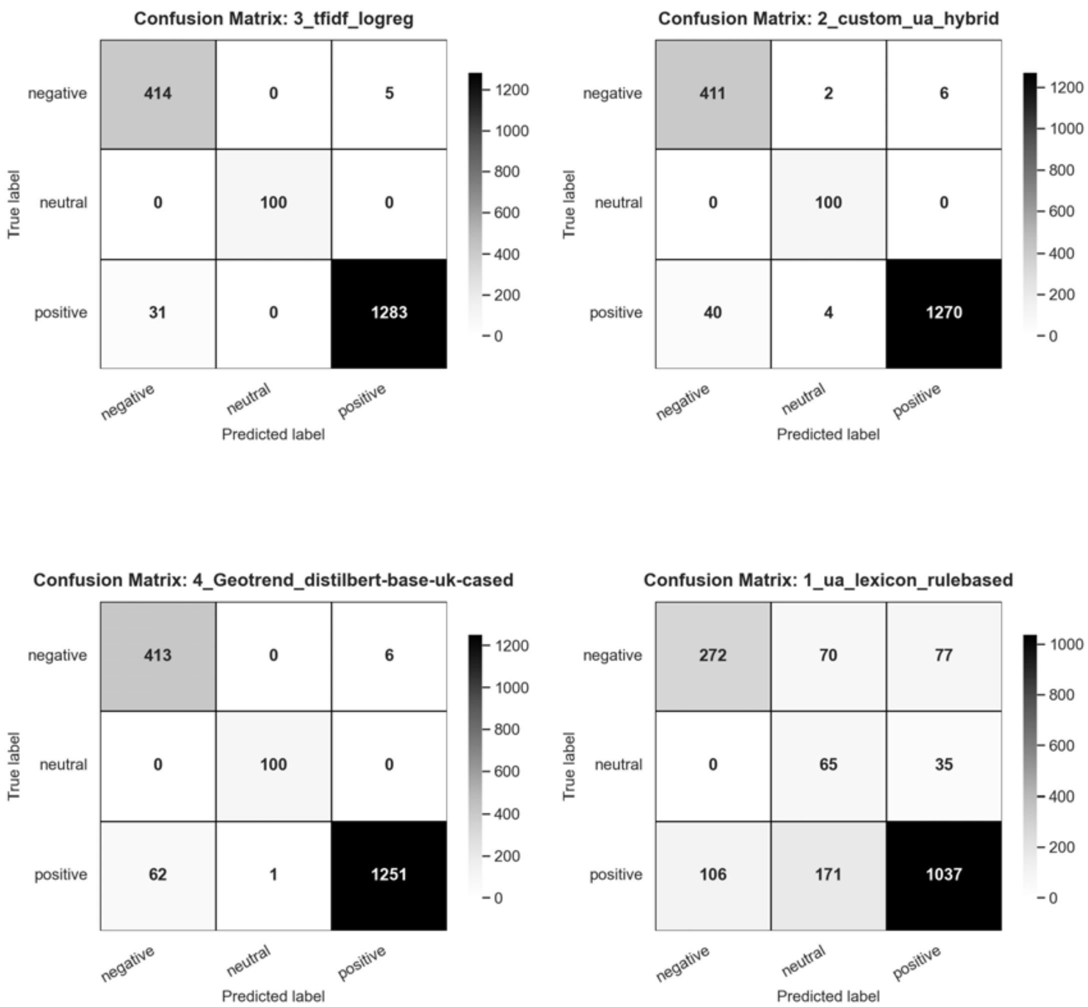

Fig. 2. Confusion matrices of sentiment analysis models

Figure 2 displays confusion matrices for the evaluated models, showing the distribution of correct and misclassified predictions across sentiment classes.

A comparison of class-level performance for the different approaches under evaluation is shown in Fig. 3. The results reveal significant differences in overall classification accuracy among the various modelling approaches. The basic lexical rule-based approach achieves an accuracy of 74.96%, highlighting the limitations of fixed-dictionary methods when processing languages with complex morphology.

Classical machine learning methods demonstrate significantly better results. The TF-IDF model combined with logistic regression achieves the highest accuracy among all the approaches studied–98.04%–demonstrating the high effectiveness of linear classifiers in sentiment analysis tasks, where polarity is often determined by explicit lexical markers.

A hybrid approach that combines lexical and statistical features achieves an accuracy of 97.16%, confirming the importance of integrating diverse feature sources to improve model robustness.

The DistilBERT-base-uk-cased transformer model also shows high performance, achieving an accuracy of 96.24%. Although its results are slightly lower than TF-IDF + Logistic Regression, transformer architectures offer more context-dependent representations of text and can better generalise to complex semantic dependencies.

Overall, the evaluation confirms that contextual modelling methods are effective for Ukrainian sentiment analysis. It also provides empirical support for the use of transformer-based techniques in practical information systems in low-resource settings.

Despite the high overall performance of several evaluated models, a detailed analysis reveals systematic challenges in sentiment classification in Ukrainian-language texts. The most frequent errors occur when sentiment polarity is weakly expressed or ambiguously conveyed in context. Specifically, neutral statements often include lexical elements that may appear sentiment-bearing outside their context, leading to misclassification by both statistical and neural models.

Another reason for classification errors arises from Ukrainian-specific linguistic characteristics. The language's morphological variability, flexible word order, and common use of intensifiers or negation constructions complicate sentiment analysis, particularly for rule-based and traditional machine learning approaches that rely mainly on surface lexical cues. Consequently, models may misinterpret polarity cues or miss important syntactic relationships affecting sentiment detection.

Transformer-based architectures show greater resilience to linguistic variations thanks to their contextual representation mechanisms. However, even these models sometimes struggle with implicitly expressed sentiment, irony, or context-dependent emotional cues that demand deeper pragmatic understanding. These limitations suggest that sentiment analysis for Ukrainian-language content remains sensitive to linguistic nuances and contextual differences.

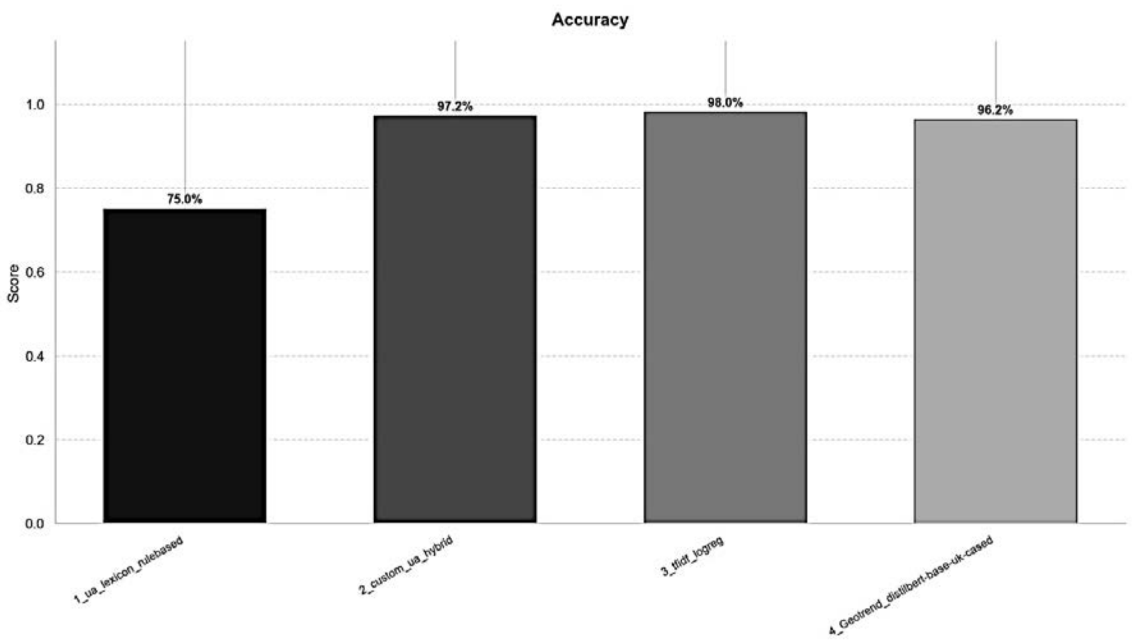

Fig. 3. Accuracy chart comparison

Overall, the results show that modern contextual language models significantly improve sentiment classification. However, additional gains are possible with hybrid modelling approaches, larger annotated datasets, and domain-specific language representations tailored to Ukrainian data.

Conclusions. This study offers a comprehensive comparison of rule-based, classical machine learning, and transformer-based approaches to Ukrainian-language sentiment analysis within a unified experimental framework. The proposed methodological design ensures consistent evaluation conditions and enables reliable comparisons across diverse modelling paradigms.

The results clearly show performance differences among the approaches and confirm that contextual representation models are effective for sentiment classification in Ukrainian texts. Transformer-based models demonstrate strong performance overall; however, classical machine learning models achieve the highest accuracy in this experimental setup. Meanwhile, classical machine learning methods achieve competitive results in controlled experiments – rule-based methods, though interpretable and resource-efficient, are less effective in complex linguistic cases.

The error analysis highlights persistent issues with linguistic ambiguity, implicit sentiment expression, and the structural variability of Ukrainian texts. These results indicate that sentiment classification accuracy remains affected by contextual and pragmatic factors, despite the use of advanced modelling techniques.

The findings support advancing Ukrainian sentiment analysis by providing a standardised evaluation framework and empirical data that endorse the application of contextual language models in real-world information systems. Future studies might aim to expand annotated corpora, improve domain adaptation strategies, and explore hybrid models that integrate linguistic expertise with contextual embeddings.

Bibliography:

1. Devlin J., Chang M.-W., Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*. Minneapolis, USA, 2019. P. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

2. Vasyliuk A., Basyuk T., Lytvyn V. Design and implementation of a Ukrainian-language educational platform for learning programming languages. *Proceedings of the 5th International Workshop on Modern Machine Learning Technologies and Data Science (MoMLeT&DS 2023).* Lviv, Ukraine, June 3, 2023. P. 406–420.

3. Basyuk T., Lomovatskyi A. Improving sentiment analysis of Ukrainian-language content by applying a rule-based approach. *CEUR Workshop Proceedings*. 2025. DOI: https://doi.org/10.31110/COLINS/2025-1/012

4. Reimers N., Gurevych I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* 2020. P. 4512–4525. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.365

5. Stigall W., Al Hafiz Khan M. A., Attota D., et al. Large Language Models Performance Comparison of Emotion and Sentiment Classification. *Proceedings of the 2024 ACM Southeast Conference (ACMSE 2024).* 2024. DOI: https://doi.org/10.1145/3603287.3651183

6. Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2302.13971.

7. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of ACL*. 2020. DOI: https://doi.org/10.18653/v1/2020.acl-main.560

8. Thapa S., Shiwakoti S., Shah S. B., et al. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*. 2025. DOI: https://doi.org/10.1007/s13278-025-01428-9

9. Singh K., Dhawan S., Bali N., et al. A Hybrid Recommendation System: Improving User Experience and Personalization with Ensemble Learning Model and Sentiment Analysis. *International Journal of Computing and Digital Systems*. 2025. DOI: https://doi.org/10.12785/ijcds/1571024952

10. Tang Z., Huang C.-Y., Li T.-C., et al. Using Contextually Aligned Online Reviews to Measure LLMs' Performance Disparities Across Language Varieties. *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2025)*. 2025. DOI: https://doi.org/10.18653/v1/2025.naacl-short.29

Ломовацький А.А., Басюк Т.М. МЕТОДИ АНАЛІЗУ НАСТРОЇВ У КОНТЕНТІ УКРАЇНСЬКОЮ МОВОЮ: ПІДХОДИ НА ОСНОВІ ПРАВИЛ, МАШИННОГО НАВЧАННЯ ТА ТРАНСФОРМЕРІВ

Аналіз тональності є важливою складовою сучасної обробки природної мови, що забезпечує автоматизоване визначення емоційного забарвлення текстових даних у різних галузях застосування, зокрема в аналізі соціальних мереж, системах підтримки прийняття рішень, маркетинговій аналітиці та *моніторингу громадської думки. Проте аналіз тональності для морфологічно складних і малоресурсних мов, зокрема української, залишається складним завданням через обмеженість розмічених корпусів, мовну варіативність, гнучкий порядок слів і складну граматичну структуру. Це ускладнює безпосереднє застосування методів, розроблених для англійської та інших високоресурсних мов.*

У статті подано систематизований огляд сучасних підходів до аналізу тональності українськомовного контенту, зосереджений на rule-based методах, класичних методах машинного навчання та трансформерних мовних моделях. Проаналізовано теоретичні засади зазначених підходів, їх переваги, обмеження та придатність до обробки українськомовних текстів в умовах обмежених ресурсів. Особливу увагу приділено rule-based методам як ефективному базовому рішенню для малоресурсних умов, моделям машинного навчання як масштабованим класифікаторам, а також трансформерним архітектурам як сучасним рішенням, що забезпечують контекстно залежні представлення тексту та здатні враховувати складні семантичні залежності.

У роботі наведено коротке порівняльне дослідження ефективності різних підходів на українськомовних даних із використанням стандартних метрик оцінювання, зокрема accuracy, precision, recall та F1-score. Отримані результати свідчать, що хоча rule-based методи залишаються ефективними в обмежених умовах і можуть використовуватися як базові моделі, трансформерні моделі демонструють вищу точність, кращу узагальнюваність і стійкість до лінгвістичної варіативності текстів. Результати дослідження сприяють систематизації підходів до аналізу тональності українськомовного контенту та окреслюють перспективи розвитку гібридних підходів у задачах обробки малоресурсних мов.

Ключові слова: *аналіз тональності, українська мова, обробка природної мови, rule-based методи, машинне навчання, трансформерні моделі, малоресурсні мови.*

Дата першого надходження статті до видання: 17.03.2026
Дата прийняття статті до друку після рецензування: 14.04.2026
Дата публікації (оприлюднення) статті 11.05.2026