JEL: C55; M31; L81

Derbentsev Vasyl
PhD in Economics
Kyiv National Economic University named after Vadym Hetman
Ukraine
ORCID: 0000-0002-8988-2526

Bezkorovainyi Vitalii
PhD in Economics
Kyiv National Economic University named after Vadym Hetman
Ukraine
ORCID: 0000-0002-4998-8385

Akhmedov Renat
Kyiv National Economic University named after Vadym Hetman
Ukraine
ORCID: 0000-0002-3084-4672

Bondarchuk Mykola
Kyiv National Economic University named after Vadym Hetman
Ukraine
ORCID: 0009-0008-7192-037X

https://doi.org/10.60022/sis.2.(02).6

EVALUATING CUSTOMER EXPERIENCE IN E-COMMERCE: MULTILINGUAL SENTIMENT ANALYSIS OF USER REVIEWS USING TRANSFORMER MODELS

Abstract. *This study addresses the challenge of multilingual sentiment analysis in e-commerce, with a focus on Ukrainian and Russian book reviews. We propose a hybrid framework based on transformer architectures that accounts for the linguistic complexity of Slavic languages and the significant class imbalance often present in customer feedback data. Using a dataset of approximately 70,000 user reviews from the Ukrainian online bookstore Yakaboo, we evaluate four model variants based on the XLM-RoBERTa architecture and compare their performance to a monolingual Ukrainian RoBERTa baseline.*

The classification task is formulated as binary: identifying low-rated reviews (1–3 stars) versus high-rated ones (4–5 stars), with the minority class comprising only 4–5% of the dataset. Our best-performing model — combining partial fine-tuning, a deep classifier, and focal loss — achieves a macro F1 score of 0.73 and an F1 score of 0.47 for the minority class, outperforming the baseline (0.66 and 0.37, respectively). The application of oversampling and focal loss proved effective in mitigating class imbalance.

Beyond technical performance, the findings underline the practical utility of accurate sentiment detection for e-commerce platforms. Effective identification of negative feedback enables companies to address customer dissatisfaction, refine product offerings, and inform personalized marketing strategies. This research contributes to advancing multilingual NLP in low-resource settings and provides a scalable solution for real-world sentiment classification in morphologically rich languages.

The scientific novelty of this research lies in the integration of multilingual transformer architectures with adaptive classifiers and class imbalance mitigation techniques, specifically tailored for real-world e-commerce review data in morphologically rich languages. The proposed framework demonstrates how domain-specific fine-tuning and architectural customization can significantly enhance sentiment classification performance in low-resource language settings.

Keywords: *Sentiment analysis, Natural Language Processing, e-commerce, customer experience, multilingual transformer models, XLM-RoBERTa, class imbalance.*

1. Introduction

In today's digital-first economy, user-generated content, particularly online reviews, has become a cornerstone of consumer interaction and business intelligence. The proliferation of e-commerce platforms, online marketplaces, and digital publishing environments has led to an exponential growth in the volume of user reviews, which — accompanied by star ratings — form a powerful dual-source of information offering both qualitative textual feedback and quantitative evaluation. According to Statista [1], the global e-commerce market reached $4.3 trillion in 2024, with strong projections for continued growth. On the one hand, this is due to the increased use of AI in digital marketing [2–4], on the other hand, this boom has placed increasing emphasis on companies' ability to monitor and respond to consumer sentiment in real time.

User-generated reviews have become especially critical for publishers, providing immediate feedback on content quality and user satisfaction. However, effectively utilizing this feedback presents substantial challenges. Online reviews are inherently unstructured and written in free-form language that often includes irony, sarcasm, slang, and ambiguity. This complexity is further compounded in Slavic languages, where lexical variation and rich morphological structures complicate text analysis [5]. Moreover, a frequent mismatch exists between the emotional tone expressed in the review text and the numeric star rating assigned by the user, complicating models that rely solely on one of these data sources to infer satisfaction [4, 6].

Sentiment analysis (SA), a key subfield of Natural Language Processing (NLP), seeks to determine the emotional polarity (positive, negative, or neutral) of a given text. With the advent of Deep Learning (DL), particularly transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) and its multilingual variants (e.g., XLM-RoBERTa) [7–8], the field has achieved significant progress. These models leverage pretraining on massive corpora to capture complex syntactic and semantic relationships, enabling more contextually aware sentiment classification. Empirical studies [9–12] have shown that transformer-based models consistently outperform conventional approaches across a wide range of NLP tasks.

Despite these advancements, two practical challenges remain underexplored. First is the issue of class imbalance in sentiment-labeled datasets. Most users are more likely to submit a review when they are highly satisfied or highly dissatisfied, resulting in a skewed distribution of sentiment labels. This imbalance can bias machine learning models, leading to poor performance in identifying underrepresented classes — typically negative reviews. Second is the inconsistent correlation between textual sentiment and star ratings, which can obscure the true emotional content of a review. These challenges are particularly pronounced in book reviews, where literary language, figurative expressions, and subjective interpretation often influence the sentiment expressed.

This paper proposes a novel approach to detecting low-rated book reviews (1 to 3 stars) by leveraging sentiment classification on multilingual (Ukrainian and Russian) data. We frame the task as a binary classification problem focused on identifying negative or neutral sentiment, regardless of the explicit star rating. Our objectives are threefold:

1. To evaluate the effectiveness of multilingual transformer-based models (specifically XLM-RoBERTa) in classifying review sentiment in a binary setting (positive vs. negative/neutral).

2. To assess the performance of a monolingual baseline model (UA-RoBERTa) trained specifically on Ukrainian data.

3. To explore the impact of class imbalance mitigation techniques, such as oversampling and focal loss, on classification performance, with particular emphasis on the underrepresented class of negative reviews (approximately 4%).

Our contributions are as follows:

- We compile and preprocess a real-world dataset of user reviews in Ukrainian and Russian.
- We compare transformer-based sentiment classification models, including monolingual and multilingual variants.
- We experiment with fine-tuning and freezing strategies for large multilingual language models.

By focusing on the sentiment classification of book reviews in a multilingual and imbalanced dataset, this research contributes to both technical and applied aspects of NLP. From a technical standpoint, it demonstrates the viability of large-scale multilingual models in low-resource and morphologically rich language settings. From an applied perspective, the insights generated may inform the development of recommender systems, editorial decision-making tools, and customer experience dashboards in publishing and e-commerce domains.

The remainder of this paper is structured as follows. Section 2, *Materials and Methods*, presents related work, dataset description, identified limitations of existing approaches, and the proposed modeling framework. Section 3, *Results and Discussion*, describes the

experimental setup, evaluates model performance, and analyzes classification results. Section 4, *Conclusions*, summarizes the key findings and outlines practical implications and directions for future research.

2. Materials and Methods

2.1. Review resent papers

Sentiment analysis (SA) has emerged as a critical tool for interpreting user-generated content in e-commerce environments [9]. As noted by Ge et al. [13], the field has evolved significantly from basic rule-based sentiment classifiers to sophisticated deep learning (DL) approaches, with transformer-based models representing the current state of the art.

Numerous studies have investigated the relationship between star ratings and review sentiment. Authors of [6] identified substantial discrepancies between numerical star ratings and textual sentiment, showing that text often conveys nuances not reflected in the rating. Bigne et al. [4] examined this misalignment in the tourism sector, while Pan and Zhang [14] were among the first to explore interactions between review text and rating valence. Subsequent studies, including [15–16] further analyzed these inconsistencies, highlighting how review length and sentiment richness may impact user interpretation and model accuracy.

Several studies have addressed SA in low-resource languages. Authors of [17] evaluated data augmentation techniques to overcome data scarcity, and paper [6] were presented a systematic review of multilingual SA methods for under-resourced languages, emphasizing transfer learning and cross-lingual modeling. Aliyu et al. [18] extended this discussion with a comprehensive overview of models, languages, and datasets applicable to low-resource SA tasks.

The rise of transformer-based architectures has fundamentally reshaped NLP research. Devlin et al. [7] introduced BERT, enabling contextualized bidirectional language representations. Conneau et al. [8] later proposed XLM-RoBERTa, a multilingual transformer trained on a large corpus of 100 languages. Barbieri et al. [19] introduced XLM-T, optimized for Twitter-based multilingual SA. Authors of [20] compared multilingual approaches on social media data, while Miah et al. [21] proposed a multimodal ensemble combining transformers and LLMs. Results of paper [22] were demonstrated the superiority of DL methods including transformer-based architectures over traditional ML in analyzing e-commerce reviews.

In a series of contributions, authors of [11, 23–25] investigated content analysis and sentiment detection in electronic media using ML techniques. Their work includes a methodology for content analysis of mass media texts [23], a study on emotional polarity classification in social networks [24], and an empirical comparison of DL models for sentiment classification in social media [25]. In a separate applied study, they analyzed social media content with ML methods to support sentiment detection pipelines in digital environments [11].

Another well-known challenge in SA is class imbalance. In real-world datasets, positive reviews often dominate, while negative reviews although typically more informative are underrepresented [26–27]. Oversampling methods such as SMOTE [28] have been widely used to address this issue by synthetically generating samples for the minority class. Additional methodological challenges include code-switching, sarcasm detection, negation handling, and cultural variation in sentiment expression.

Several review papers have synthesized the evolving landscape of SA. Authors of [9, 29] analyzed the application of sentiment analysis on e-commerce platforms, while Mao et al. [30] provided a comprehensive review of sentiment methods, challenges, and trends in both traditional ML and modern DL contexts. Results of paper [31] emphasized the importance of feature engineering for Chinese reviews. Authors of [10] surveyed multilingual DL approaches for sentiment analysis in social media.

Research focusing on Slavic languages (especially for Ukrainian) remains relatively limited. Thus, Radchenko [32] proposed fine-tuned RoBERTa model for Ukrainian language, and Prytula [5] benchmarked transformer models including BERT, DistilBERT, XLM-RoBERTa, and UKR-RoBERTa on Ukrainian sentiment classification.

In summary, recent literature demonstrates the rapid advancement of sentiment analysis techniques, particularly in multilingual and low-resource contexts. However, limited research has explored sentiment classification in Slavic languages using deep transformer-based models in imbalanced datasets. Our work builds on this gap by applying multilingual transformers to Ukrainian and Russian reviews and addressing class imbalance through oversampling and loss-based optimization.

2.2. Limitations of Existing Approaches

Despite the availability of various sentiment analysis tools and pretrained models, several critical limitations emerge when applying them to the task of analyzing multilingual user reviews, particularly for low-resource and morphologically rich languages such as Ukrainian and Russian. These limitations can be categorized into five key areas:

1. Insufficient Language Support. Most widely used sentiment analysis libraries, such as *TextBlob, VADER*, and *Afinn*, are primarily optimized for English-language content. Their performance deteriorates significantly when applied to Slavic languages, which differ substantially in morphology, syntax, and semantics. Furthermore, these libraries typically lack native support for Ukrainian and Russian, leading to unreliable outputs and limited generalization [5, 12, 18].

2. Constraints of Proprietary Large Language Models (LLMs). While state-of-the-art commercial models (e.g., OpenAI GPT, Google Gemini, Anthropic Claude) demonstrate strong multilingual capabilities, they are accessible only via API-based interfaces that

often involve high usage costs. This poses challenges for large-scale academic studies, particularly those requiring privacy, reproducibility, or transparency. Additionally, their black-box nature limits control over internal processes and domain-specific adaptations [21].

3. Limited Customization in Cloud-Based Sentiment Platforms. Cloud-based NLP platforms such as Google Cloud Natural Language API or IBM Watson offer sentiment analysis as part of broader content analytics services. However, these tools are often subscription-based and offer limited ability to fine-tune models for specific domains. Their general-purpose pretrained models tend to oversimplify domain-specific sentiment, especially in contexts such as literary critique or user reviews in e-commerce [25].

4. Translation-Based Workarounds. A common workaround involves translating user reviews into English and then applying established English-language sentiment models. While convenient, this approach introduces additional sources of error. Machine translation may distort sentiment-laden expressions, weaken contextual nuance, or omit subtle emotional cues — particularly problematic for reviews that rely on idioms, sarcasm, or domain-specific phrasing. As a result, the final sentiment prediction may deviate significantly from the original intent [12, 31].

5. Misalignment Between Text and Rating Labels. In real-world datasets, star ratings and textual sentiment are frequently misaligned. For instance, a user might assign a high star rating while expressing critical opinions in the review text or vice versa. This misalignment undermines the validity of models trained exclusively on rating labels, as they fail to capture mixed or nuanced sentiment expressions that are common in subjective review genres such as book evaluations [4, 6, 24].

Taken together, these limitations underscore the need for domain-adapted and language-specific sentiment models trained on real multilingual data. In this study, we aim to overcome these barriers by applying fine-tuned transformer-based architectures and custom classification components, tailored to the linguistic and structural characteristics of Ukrainian and Russian user reviews.

2.3. Dataset Description

This study utilizes the Yakaboo Book Reviews dataset, a large-scale corpus of user-generated reviews collected from the Ukrainian online bookstore Yakaboo [33]. The dataset comprises 69,256 annotated entries, each containing a review text, a star rating (from 1 to 5), and associated metadata, such as publication date, book ID, and review ID. The reviews were published between May 2014 and August 2020, with the median date falling in April 2019.

To frame the sentiment classification task, we restructured the original 5-point star rating system into a binary classification problem. Reviews with ratings from 1 to 3 stars were grouped into the Low (1–3) class (label 1), while 4 and 5-star reviews were grouped into the High (4–5) class (label 0). This framing emphasizes the identification of less favorable feedback, which is both underrepresented and business-critical. The dataset is highly imbalanced, with approximately 83% of reviews rated 5 stars, and only 4–5% rated from 1 to 3 stars (see Figure 1).

Linguistically, the corpus is nearly evenly split between Ukrainian (49.5%) and Russian (51.5%), making it suitable for evaluating multilingual language models. On average, *Low-rated* reviews tend to be longer than *High-rated* ones in both languages (e.g., Russian: 194 vs. 179 words; Ukrainian: 217 vs. 197 words), suggesting that users may express dissatisfaction more elaborately (see Figure 2).

In our analysis, we used a concatenated version of both the review title and the review body as the

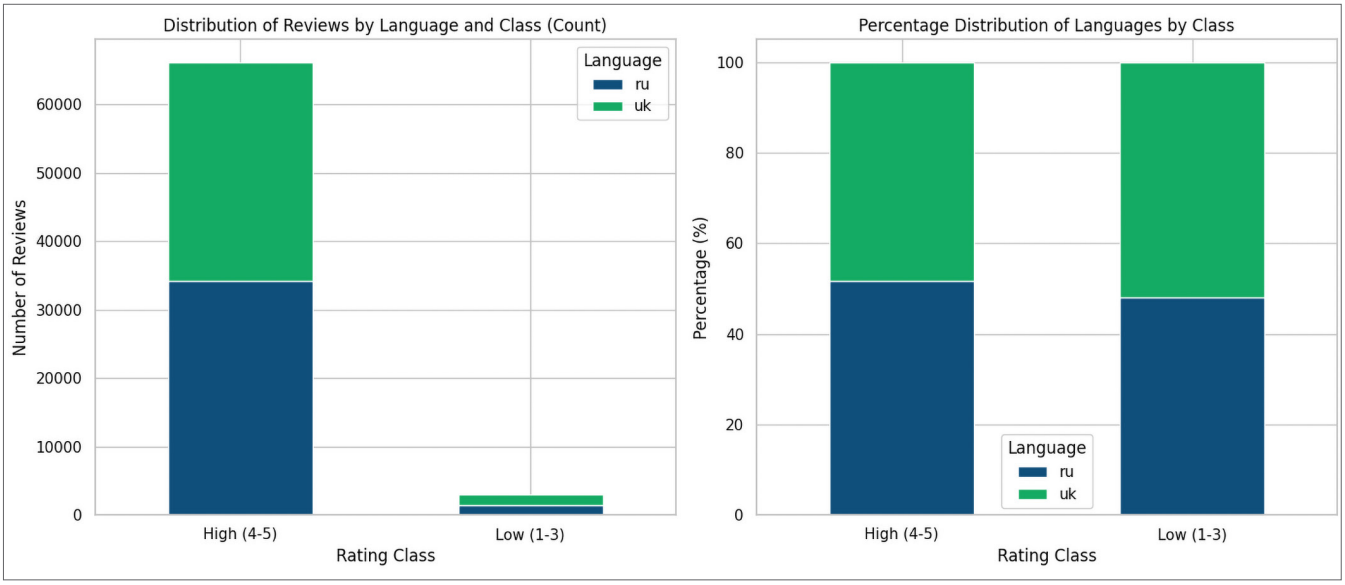


Figure 1. Distribution of Reviews by Language and Class

Source: calculated and designed by the authors

input text. To ensure consistency and improve model performance, we applied several preprocessing steps:
- Missing values in both review Title and review fields were replaced with empty strings.
- Excess whitespace and special characters were normalized using regular expressions.
- The final input text (text) was constructed by concatenating the review title and review fields with a space separator, followed by additional cleaning to remove redundant spaces.
- Reviews with empty or invalid text after preprocessing were filtered out.
- All text was lowercased to reduce vocabulary size and improve tokenizer efficiency.
- Potential HTML tags were stripped using regex-based patterns.
- Redundant punctuation (e.g., "!!!" or "???") was normalized to a single instance to retain emotional intensity while avoiding overfitting.

This preprocessing pipeline ensured cleaner and more semantically meaningful inputs, suitable for tokenization by transformer-based language models.

2.4. Proposal Approach

To address the limitations of existing multilingual sentiment analysis tools, we propose a hybrid framework that combines fine-tuned transformer architectures with custom neural classifiers, optimized specifically for imbalanced, multilingual book reviews in Ukrainian and Russian. Our approach prioritizes:
- language-specific adaptability, addressing the morphological richness of Slavic languages;
- class imbalance mitigation, through techniques such as oversampling and focal loss;
- computational efficiency, via partial fine-tuning and frozen embeddings.

The main objective of this study is to predict review ratings (on a 1–5 star scale) using only textual content, with a particular focus on detecting low-rated (1–3 stars) reviews. We frame this as a binary sentiment classification task, targeting negative or neutral sentiment regardless of the explicit star rating. This task presents several challenges, including:
- significant class imbalance;
- subtle linguistic cues;
- the need to capture context-dependent sentiment patterns across two languages.

To address these issues, we implement a multilayer modeling framework that integrates pretrained multilingual transformers with task-specific classifiers.

Transformer-based language models such as BERT, RoBERTa, and XLM-RoBERTa have transformed NLP by effectively capturing semantic and syntactic nuances using self-attention mechanisms [7–8]. These models are pretrained on massive multilingual corpora, enabling strong cross-lingual understanding and robust performance across languages. In particular, XLM-RoBERTa is well suited for our task, as it supports morphologically rich languages and handles informal, user-generated content effectively.

Unlike monolingual models trained on English or Ukrainian, XLM-RoBERTa's cross-lingual capacity allows us to use a single model for both Ukrainian and Russian reviews, thereby reducing the need for separate pipelines. Moreover, its pretrained embeddings allow downstream classifiers to better interpret idioms, code-switching, and informal sentiment expressions.

We designed and tested four architectural variants (Table 1), each exploring different trade-offs in model capacity, fine-tuning depth, and class imbalance robustness:

UA RoBERTa. A monolingual model pretrained on Ukrainian corpora [32], fine-tuned on the top two

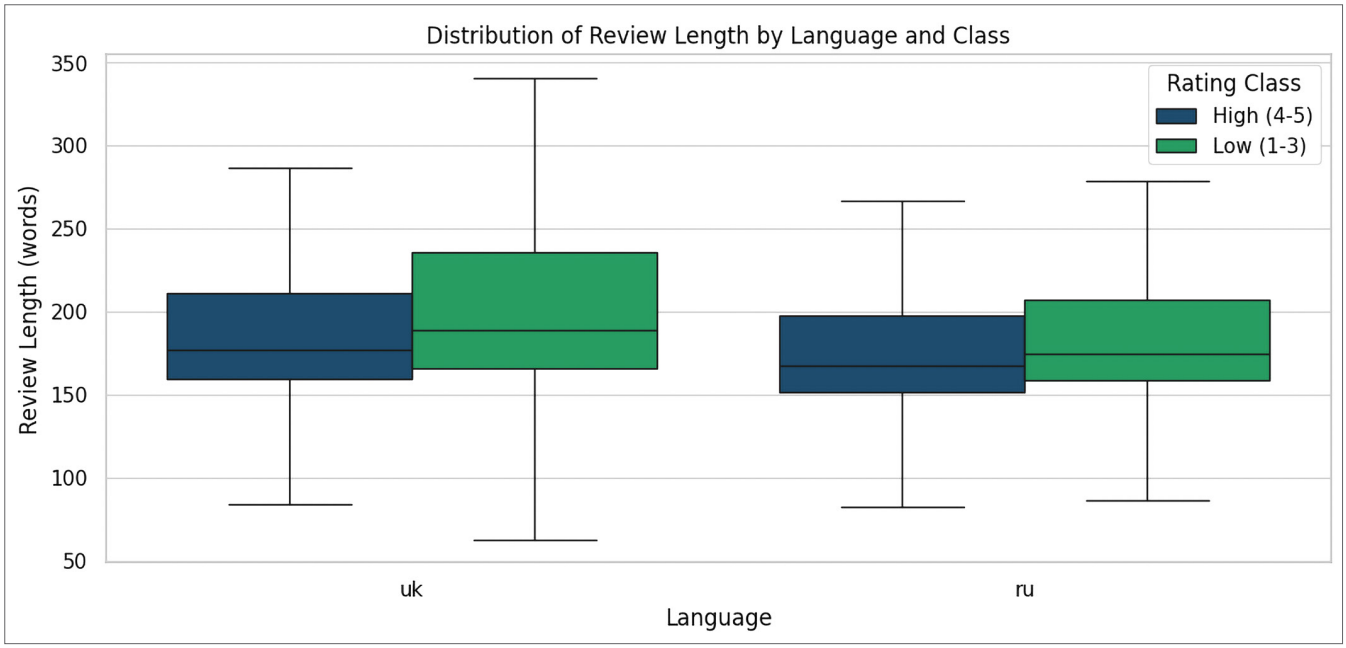


Figure 2. Distribution of Review's Length by Language and Class
Source: calculated and designed by the authors

Table 1

Architectural Variants

Variant	Backbone	Fine-Tuning	Classifier	Key Innovation
Baseline	UA-RoBERTa	Top 2 layers	Dense layer	Monolingual benchmark tailored for Ukrainian text
Variant 1	XLM-RoBERTa	Top 2 layers	Dense layer	Lightweight multilingual adaptation with reduced training cost
Variant 2	XLM-RoBERTa (frozen)	None	BiLSTM / CNN	Leverages frozen multilingual embeddings and adds sequential feature extraction
Variant 3	XLM-RoBERTa	Top 2 layers	BiLSTM/CNN + Focal Loss	Integrates partial fine-tuning, advanced classifier, and imbalance-aware loss

Source: designed by the authors

transformer layers and equipped with a dense classification head. Serves as a benchmark to compare against multilingual alternatives.

Fine-tuned XLM-RoBERTa with Dense Classifier. his configuration partially fine-tunes the top two layers of XLM-RoBERTa and applies a simple dense output layer (typically based on the [CLS] token) to map embeddings to sentiment classes. It offers a balance between performance and training cost.

Frozen XLM-RoBERTa with Custom BiLSTM/CNN Classifier. In this setup, the entire XLM-RoBERTa model is frozen, and its output embeddings are passed to a custom classifier composed of:

- BiLSTM: to model long-range dependencies and sequential sentiment patterns;
- 1D CNN: to extract n-gram features and localized emotional signals.

These classifiers are evaluated separately to identify which structure better utilizes the transformer-generated representations.

Fine-tuned XLM-RoBERTa with BiLSTM/CNN Classifier. This advanced variant integrates the strengths of both Variant 1 and Variant 2. It partially fine-tunes the top two layers of XLM-RoBERTa while adding a deep classifier (BiLSTM or CNN) and applying **focal loss** to dynamically adjust learning based on class distribution. This configuration directly targets the minority class problem and enhances detection of subtle or mixed sentiment expressions.

These variants represent a spectrum of trade-offs in terms of language specialization, computational efficiency, and classification complexity:

- The UA-RoBERTa baseline is strong for Ukrainian text but lacks multilingual capacity.
- Variant 1 provides an efficient and scalable multilingual alternative with partial fine-tuning.
- Variant 2 eliminates fine-tuning altogether, relying solely on pretrained embeddings and downstream modelling — suitable for low-resource scenarios.
- Variant 3 integrates fine-tuning, advanced classification, and class imbalance optimization, offering the most robust performance for imbalanced multilingual sentiment classification.

This modeling framework allows for a comparative evaluation of how different architectural choices affect classification accuracy, particularly for detecting negative or neutral user feedback in real-world, multilingual e-commerce environments.

3. Results and Discussion

3.1. Experimental Setup

To ensure comparability across all experiments, we standardized the training and evaluation procedures for each model variant. All models were trained using the Adam optimizer with an initial learning rate of 2e-5 and a linear learning rate scheduler with warm-up steps [34]. Training was conducted for a maximum of 5 epochs, with early stopping based on validation loss to prevent overfitting. The batch size was set to 16.

Considering the average length of user reviews (see Figure 2), we set a maximum input sequence length of 256 tokens for all models to ensure consistency in input representation across architectures.

Two loss functions were evaluated:

- **Binary Cross-Entropy (BCE)**: the standard choice for binary classification tasks;
- **Focal Loss:** employed to address class imbalance by dynamically scaling the contribution of each sample based on prediction confidence, emphasizing harder-to-classify examples [35].

To further mitigate the effects of class imbalance, the following techniques were applied [26]:

- **Class weighting**: weights inversely proportional to class frequency were applied during training;
- **Random oversampling**: minority class instances were oversampled using RandomOverSampler from the *imbalanced-learn* library. This was applied only to the training set to prevent data leakage.

Each of the four model variants based on XLM-RoBERTa was evaluated on the binary sentiment classification task, distinguishing between low-rated reviews (1–3 stars) and high-rated reviews (4–5 stars). Given that only about 4% of the dataset corresponds to low-rated reviews, the task presents a significant imbalance challenge.

To ensure robust evaluation, we used an 80/20 stratified train-test split of the dataset, preserving the proportion of the minority class in both subsets.

All experiments were implemented using the PyTorch DL framework and Hugging Face Transformers

[36], and executed in Google Colab Pro+, leveraging NVIDIA L4 GPUs (24 GB VRAM). Mixed precision training (FP16) was enabled to accelerate training and reduce memory usage.

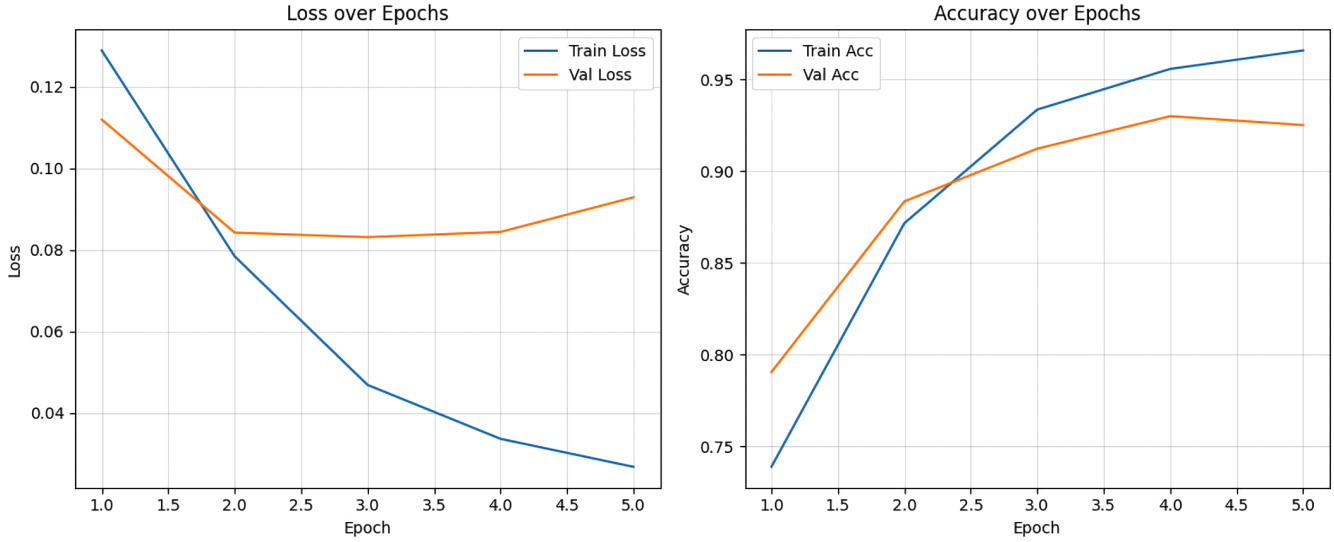
Figure 3 illustrates the training and validation loss and accuracy curves for the Baseline model (a) and Variant 3 (b). The Baseline model shows moderate overfitting: while training loss continues to decrease steadily, validation loss plateaus early and slightly increases after the third epoch. Accuracy follows a similar pattern, with a growing gap between training and validation performance. In contrast, Variant 3 demonstrates smoother convergence, with both loss and accuracy curves stabilizing by the fourth epoch. The smaller gap between training and validation metrics indicates better generalization, confirming the benefits of partial fine-tuning, focal loss, and the deeper classifier used in this configuration.

3.2. Classification Performance

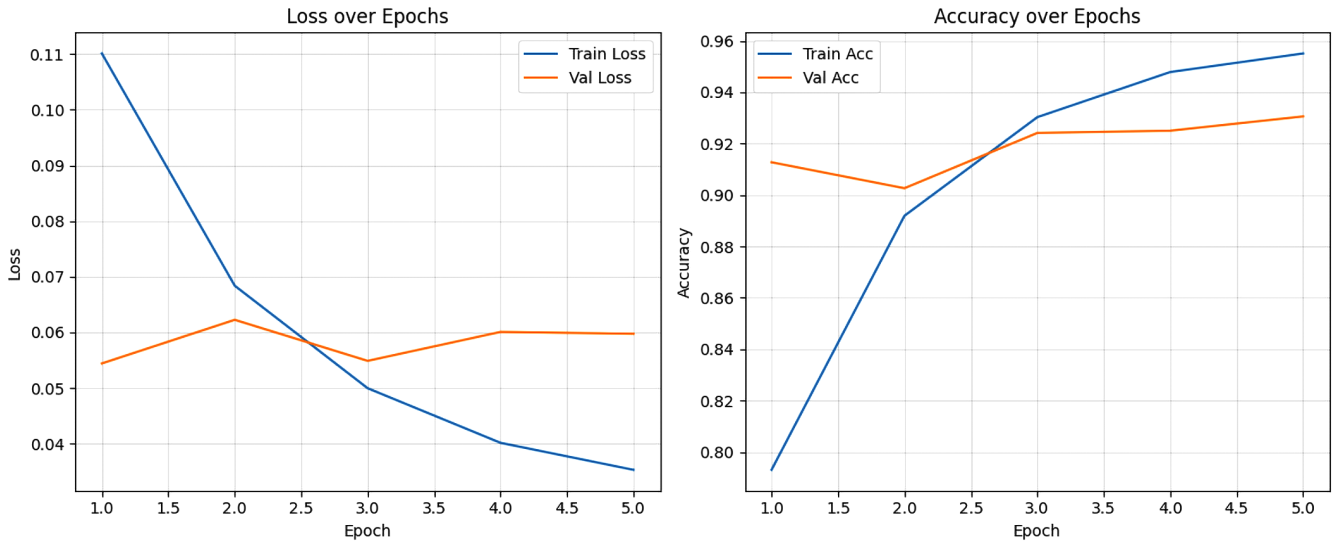
To evaluate model performance, we used a comprehensive set of metrics that reflect both overall accuracy and the models' ability to detect the minority class. Given the imbalanced nature of the dataset, particular attention was paid to performance on low-rated reviews (1–3 stars).

The following evaluation metrics were employed:

- **Macro F1**: the unweighted average of F1 scores across both classes, providing a balanced measure of performance regardless of class distribution;
- **F1 (Low)**: F1 score for the minority "Low-rated" class; a harmonic mean of precision and recall, this metric reflects the model's ability to detect negative/neutral sentiment;
- **Precision (Low)**: proportion of reviews predicted as "Low-rated" that were actually correct;
- **Recall (Low)**: proportion of actual "Low-rated" reviews that were correctly identified;



(a)



(b)

Figure 3. Training and validation loss/accuracy curves for the Baseline (a) and Variant 3 (b) models

Source: calculated and designed by the authors

Table 2

Classification Performance by Model Variant

Variant	Macro F1	F1 (Low)	Precision (Low)	Recall (Low)	ROC AUC	Inference Time (ms)	Accuracy
Baseline	0.66	0.37	0.28	0.54	0.76	32	0.92
Variant 1	0.7	0.43	0.31	**0.71**	0.83	36	**0.93**
Variant 2	0.61	0.27	0.29	0.25	0.75	**21**	0.90
Variant 3	**0.73**	**0.47**	**0.39**	0.6	**0.84**	41	0.92

Source: calculated by the authors

- **ROC AUC**: the area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between positive and negative sentiment across all thresholds;
- **Inference Time (ms)**: average time required to classify a single review; this metric is important for assessing real-world deployment feasibility;
- **Accuracy**: overall proportion of correctly classified reviews, though less informative in imbalanced settings.

These metrics enable a holistic comparison of model performance, especially in terms of the ability to capture underrepresented sentiment classes. Table 2 summarizes the classification results for all four

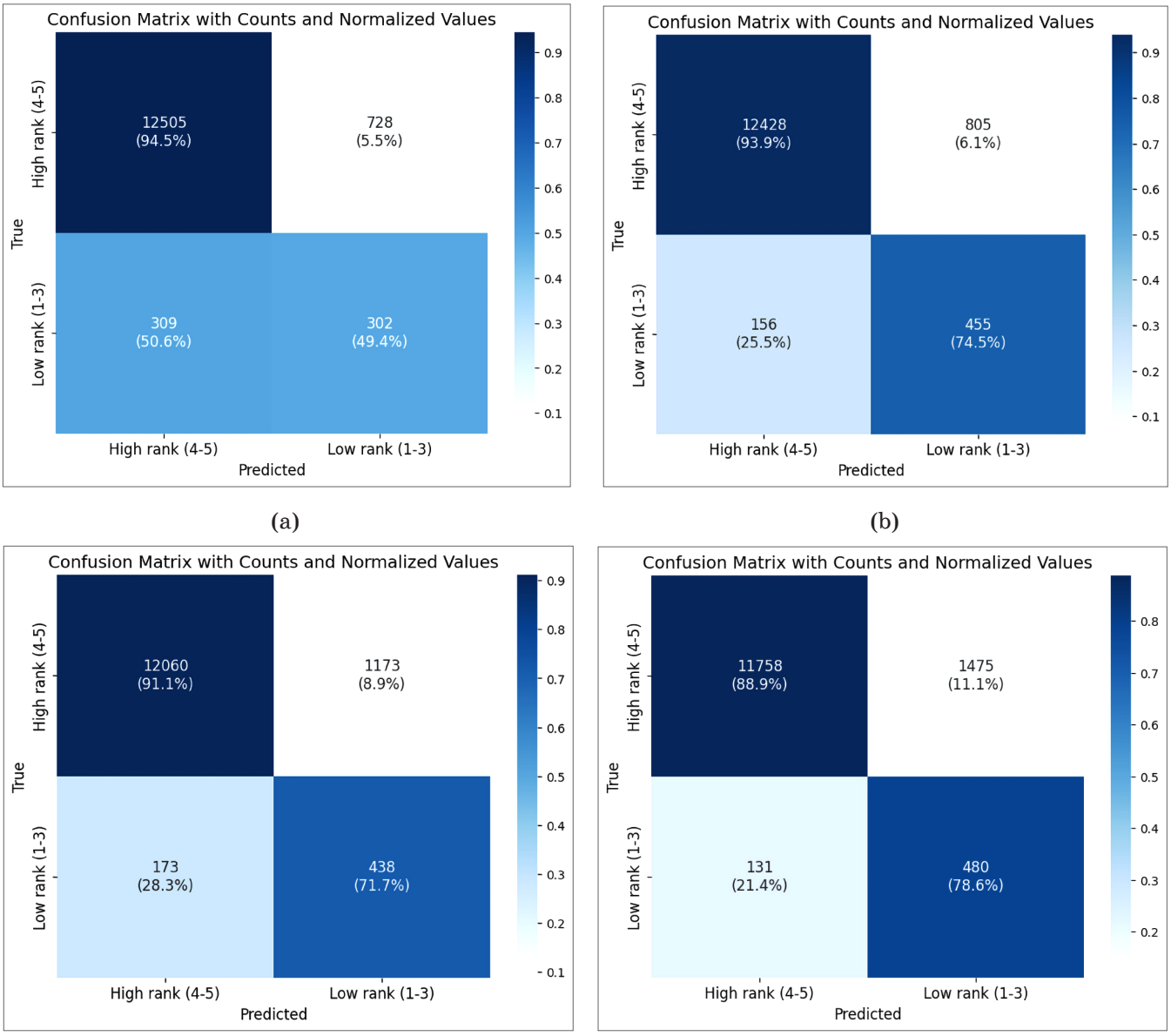


Figure 4. Confusion Matrixes for Baseline (a), Variant 1 (b), Variant 2 (c), and Variant 3 (d) model's architectures

Source: calculated and designed by the authors

model variants tested on the binary sentiment classification task.

The results clearly demonstrate that multilingual transformer models significantly outperform the monolingual baseline in identifying low-rated reviews. Variant 3, which combines partial fine-tuning, an advanced BiLSTM/CNN classifier, and focal loss, achieved the highest Macro F1, F1 (Low), and ROC AUC, indicating its superior ability to detect negative and neutral sentiment in imbalanced data.

Variant 1, which applies a simpler dense classification head, also performed well — particularly in recall, suggesting a strong ability to detect most low-rated cases, albeit with somewhat lower precision. In contrast, Variant 2, which uses frozen embeddings with no fine-tuning, showed the fastest inference time but had the weakest performance across all key metrics, especially on the minority class.

The Baseline model achieved high overall accuracy but exhibited a strong bias toward the dominant class (high-rated reviews), as evidenced by its lower recall and F1 score for the minority class.

These findings validate the effectiveness of using focal loss and partial fine-tuning in improving performance on underrepresented classes, and highlight the importance of classifier architecture in leveraging transformer embeddings effectively.

In addition to quantitative metrics, we also examined Confusion Matrices (Figure 4) to gain deeper insights into how each model handles class imbalance.

- The Baseline model (a) shows a strong bias toward the majority "High-rated" class, with a relatively low number of correctly identified "Low-rated" reviews and many false negatives.
- Variant 1 (b) demonstrates clear improvement, correctly classifying more Low-rated reviews than the baseline, though at the cost of some false positives (i.e., High-rated reviews misclassified as Low-rated).
- Variant 2 (c) exhibits the weakest performance in detecting Low-rated reviews, with a high number of misclassifications and low recall, confirming the limitations of using frozen embeddings without fine-tuning.
- Variant 3 (d) achieves the highest number of correctly identified Low-rated reviews, highlighting its strength in addressing class imbalance. However, it also shows an increased number of High-rated reviews misclassified as Low-rated, indicating a trade-off between recall and precision.

These visualizations support the quantitative findings presented in Table 2 and emphasize the importance of balancing recall and precision when dealing with minority sentiment classes.

To gain deeper insights into model behavior, we conducted a qualitative analysis of misclassified reviews. This analysis revealed several recurring patterns:

1. Mixed sentiment expressions: All models struggled with reviews that simultaneously contained both positive and negative opinions. For instance, reviews praising the book's plot but criticizing its translation or vice versa were frequently misclassified due to the presence of conflicting sentiment cues.

2. Subtle negation and irony: Reviews with nuanced linguistic structures, especially involving negation or **ironic phrasing**, posed significant challenges. A typical example is the phrase *"не можу сказати, що книга погана"* ("I cannot say the book is bad"), which was sometimes misinterpreted as negative, despite conveying overall positive sentiment.

3. Rating-text inconsistency: A notable proportion of misclassifications occurred in cases where the star rating and textual sentiment were not aligned. For example, reviews written in predominantly positive language but assigned a low rating (or the opposite) often confused the models, highlighting the limitations of using ratings as ground truth sentiment labels.

4. Code-switching and multilingual noise: Reviews containing a mixture of Ukrainian and Russian, or incorporating English words and phrases, exhibited higher error rates — especially for the Baseline model. In contrast, Variants 1 and 3, powered by XLM-RoBERTa, demonstrated greater robustness to such linguistic variation due to their multilingual pretraining.

These observations reinforce the importance of handling ambiguous sentiment, multilingual input, and label inconsistency in real-world sentiment analysis applications. They also suggest potential avenues for further improvement, such as incorporating irony detection mechanisms, attention-based alignment with ratings, or code-switching aware tokenization strategies.

4. Conclusions

This study addressed the challenge of sentiment classification in imbalanced, multilingual user reviews, focusing on the identification of low-rated book reviews written in Ukrainian and Russian. We proposed and evaluated four architectural variants based on transformer-based models, including multilingual and monolingual backbones, custom classifier components, and class imbalance mitigation techniques.

Our results demonstrate that multilingual transformer models, particularly XLM-RoBERTa, substantially outperform a monolingual baseline (UA-RoBERTa) in detecting low-rated reviews. Among all configurations, Variant 3 — which combines partial fine-tuning, a deep BiLSTM/CNN classifier, and focal loss — achieved the best overall results, with an F1 score of 0.47 for the minority class and a macro F1 of 0.73, representing a substantial improvement over the baseline. The model also achieved a strong ROC AUC of 0.84, indicating high discriminative ability.

In contrast, models relying on frozen embeddings (Variant 2) underperformed in all metrics related to minority class detection, despite offering the fastest inference. This highlights the importance of targeted fine-tuning even when working with powerful pretrained architectures. Variant 1 demonstrated a good balance between recall and speed, showing that

lightweight solutions can remain competitive when architectural choices are well-optimized.

Beyond numerical performance, qualitative error analysis revealed important challenges that remain open: handling mixed sentiment, interpreting negation and irony, resolving text-rating inconsistencies, and processing code-switched input. These aspects point to the inherent complexity of real-world sentiment analysis in multilingual and user-generated settings.

From a practical standpoint, the proposed framework offers a scalable approach for automated sentiment monitoring in digital publishing, e-commerce platforms, and recommendation systems, particularly in low-resource linguistic environments. It demonstrates how multilingual pretrained models can be adapted to underrepresented languages and real-world tasks through focused fine-tuning and classifier design.

Future work may focus on the integration of irony detection, multitask learning for sentiment and rating alignment, and the use of code-switching-aware pretraining to further enhance model robustness. Additionally, further evaluation on external datasets and different domains (e.g., product reviews, news comments) will help assess the generalizability of the proposed approach.

Overall, this research contributes to advancing multilingual sentiment analysis in underrepresented language settings, and provides a foundation for building more context-aware, fair, and accurate review classification systems.

5. References

[1] Statista. (2024). E-commerce worldwide — Statista Digital Market Outlook 2024. https://statista.com/outlook/dmo/ecommerce/worldwide.

[2] Turlakova, S.S., & Shumilo, Ya.M. (2025). Infl uence of AI Tools on Consumer Behavior Ma-nagement in Digital Marketing. Sci. innov., 21(1), 67–81. https://doi.org/10.15407/scine21.01.067.

[3] Oleksiuk, O., & Shafalyuk, O. (2023). Management of pharmaceutical online retail through a regional marketplace with neural network and statistical analytical tools. Neuro-Fuzzy Modeling Techniques in Economics, 12, 155–174. http://doi.org/10.33111/nfmte.2023.155.

[4] Bigne, E., Ruiz, C., Perez-Cabañero, C., et al. (2023). Are customer star ratings and sentiments aligned? A deep learning study of the customer service experience in tourism destinations. *Service Business*, *17*, 281–314. https://doi.org/10.1007/s11628–023–00524–0.

[5] Prytula, M. (2024). Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence*, *2*, 85–98. https://doi.org/10.15407/jai2024.02.085.

[6] Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management*, *54*, Article 102132. https://doi.org/10.1016/j.ijinfomgt.2020.102132.

[7] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). 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: Human Language Technologies*, 4171–4186. https://doi.org/10.18653/v1/N19–1423.

[8] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.

[9] Huang, H., Asemi Zavareh, A., & Mustafa, M.B. (2023). Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. *IEEE Access*, *11*, 93141–93162. https://doi.org/10.1109/ACCESS.2023.3307308.

[10] Agüero-Torales, M.M., Abreu Salas, J.I., & López-Herrera, A.G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing*, *107*, Article 107373. https://doi.org/10.1016/j.asoc.2021.107373.

[11] Derbentsev, V.D., Bezkorovainyi, V.S., Matviychuk, A.V., Pomazun, O.M., Hrabariev, A.V., & Hostryk, A.M. (2023). A comparative study of deep learning models for sentiment analysis of social media texts. *CEUR Workshop Proceedings, 3465,* 168–188. https://ceur-ws.org/Vol-3465/paper18.pdf.

[12] Mabokela, K.R., Celik, T., & Raborife, M. (2024). Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access*, *12*, 13100–13122. https://doi.org/10.1109/ACCESS.2022.3224136.

[13] Ge, J., Alonso-Vazquez, M., & Gretzel, U. (2018). Sentiment analysis: a review. *Advances in social media for travel, tourism and hospitality: new perspectives, practice and cases.* 243–261. https://doi.org/10.4324/9781315565736.

[14] Pan, Y., & Zhang, J.Q. (2009). Is a "star" worth a thousand words? The interplay between product-review texts and rating valences. *European Journal of Marketing*, *43*(11/12), 1269–1280. https://doi.org/10.1108/03090560910989876.

[15] Cho, H., Hasija, S., & Sosa, M. (2021). *Reading Between the Stars: Understanding the Effects of Online Customer Reviews on Product Demand. INSEAD Working Paper No. 2021/07/TOM.* https://doi.org/10.2139/ssrn.3240453.

[16] Altab, H.M., Yinping, M., & Adu-Yeboah, S.S. (2022). Understanding Online Consumer Textual Reviews and Rating: Review Length With Moderated Multiple Regression Analysis Approach. *SAGE Open*, *12*(2). https://doi.org/10.1177/21582440221104806.

[17] Thakkar, G., Mikelić Preradović, N., & Tadić, M. (2024). Examining Sentiment Analysis for Low-Resource Languages with Data Augmentation Techniques. *Eng*, *5*(4), 2920–2942. https://doi.org/10.3390/eng5040152.

[18] Aliyu, Y., Sarlan, A., Danyaro, K.U., B.A. Rahman, A.S., & Abdullahi, M. (2024). Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access*, *12*, 10300–10323. https://doi.org/10.1109/ACCESS.2024.3398635.

[19] Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of the 13th Conference on Language Resources and Evaluation, LREC 2022,* 258–266. https://aclanthology.org/2022.lrec-1.27

[20] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., et al. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, *35*(29), 1–17. https://doi.org/10.1007/s00521-023-08629-3.

[21] Miah, M.S.U., Kabir, M.M., Sarwar, T.B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, *14*, 19603. https://doi.org/10.1038/s41598-024-60210-7.

[22] Desai, T., & Meva, D. (2023). Comparative Analysis of Different Machine Learning Approaches for Sentiment Analysis. In H. Sharma, V. Shrivastava, K.K. Bharti, & L. Wang (Eds.), *Lecture Notes in Networks and Systems: Vol. 686. Communication and Intelligent Systems* (pp. 175–185). Springer, Singapore. https://doi.org/10.1007/978-981-99-2100-3_15.

[23] Matviychuk, A., Derbentsev, V., Bezkorovainyi, V., Kmytiuk, T., & Hostryk, A. (2024). Leveraging Artificial Intelligence and Large Language Models for Fake Content Detection in Digital Media. *CEUR Workshop Proceedings, 3933*, 75–92. https://ceur-ws.org/Vol-3933/Paper_7.pdf.

[24] Akhmedov, R.R., Bezkorovainyi, V.S., & Danylchenko, T.V. (2021). Methodology of content analysis of electronic mass media. *Ekonomichnyi Prostir (Economic Scope)*, *176*, 141–145. https://doi.org/10.32782/2224-6282/176-25.

[25] Derbentsev, V., Bezkorovainyi, V., & Akhmedov, R. (2020). Machine learning approach of analysis of emotional polarity of electronic social media. *Neuro-Fuzzy Modeling Techniques in Economics, 9,* 95–137. http://doi.org/10.33111/nfmte.2020.095.

[26] Lemaitre, G., Nogueira, F., & Aridas, C.K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(1), 559–563. https://doi.org/10.48550/arXiv.1609.06570.

[27] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

[28] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953.

[29] Firza, N., Bakiu, A., & Monaco, A. (2025). Machine Learning for Quality Diagnostics: Insights into Consumer Electronics Evaluation. *Electronics*, *14*(5), Article 939. https://doi.org/10.3390/electronics14050939.

[30] Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University — Computer and Information Sciences*, *36*(4), Article 102048. https://doi.org/10.1016/j.jksuci.2024.102048.

[31] Zheng, L., Wang, H., & Gao, S. (2018). Sentimental feature selection for sentiment analysis of Chinese online reviews. *International Journal of Machine Learning and Cybernetics*, *9*(1), 75–84. https://doi.org/10.1007/s13042-015-0347-4.

[32] Radchenko, V. (2020). *ukr-roberta-base* [Model]. *Hugging Face.* https://huggingface.co/youscan/ukr-roberta-base

[33] Yakaboo Book Reviews. (2020) *Curated list of Ukrainian natural language processing (NLP) resources (corpora, pretrained models, libraries, etc.)*. https://github.com/osyvokon/awesome-ukrainian-nlp

[34] Kingma, D.P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1412.6980.

[35] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2980–2988). IEEE. https://doi.org/10.1109/ICCV.2017.324.

[36] Hugging Face Team. (2023). *Transformers* [Software library]. *Hugging Face.* https://huggingface.co/transformers/

Received: 05/05/2024
Accepted: 12/09/2024
Published: 17.12.2024

Дербенцев Василь
кандидат економічних наук,
Київський національний економічний університет імені Вадима Гетьмана,
Україна
ORCID: 0000-0002-8988-2526

Безкоровайний Віталій
кандидат економічних наук,
Київський національний економічний університет імені Вадима Гетьмана,
Україна
ORCID: 0000-0002-4998-8385

Ахмедов Ренат
Київський національний економічний університет імені Вадима Гетьмана,
Україна
ORCID: 0000-0002-3084-4672

Бондарчук Микола
Київський національний економічний університет імені Вадима Гетьмана,
Україна
ORCID: 0009-0008-7192-037X

https://doi.org/10.60022/sis.2.(02).6

ОЦІНКА КЛІЄНТСЬКОГО ДОСВІДУ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ: БАГАТОМОВНИЙ СЕНТИМЕНТ АНАЛІЗ ВІДГУКІВ КОРИСТУВАЧІВ З ВИКОРИСТАННЯМ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ

Анотація. *У роботі розглянуто задачу багатомовного аналізу тональності в середовищі електронної комерції на прикладі рецензій українською та російською мовами на книги. Запропоновано гібридний підхід, що базується на трансформерних архітектурах і враховує як мовну складність слов'янських мов, так і суттєвий дисбаланс класів у даних користувацьких відгуків. Для дослідження використано корпус з приблизно 70 000 рецензій, зібраних на українській онлайн-платформі Yakaboo.*

Класифікаційна задача сформульована як бінарна: виявлення відгуків із низьким рейтингом (1–3 зірки) проти високих (4–5 зірок), причому частка міноритарного класу становить лише 4–5%. Найкраща модель — частково донавчена XLM-RoBERTa з глибоким класифікатором і функцією втрат focal loss — досягла макро-F1 = 0.73 і F1 для міноритарного класу = 0.47, що суттєво перевищує результати базової мономовної моделі (0.66 і 0.37 відповідно). Застосування «передискретизації більшості» (oversampling) та focal loss довело свою ефективність у подоланні дисбалансу.

Результати дослідження засвідчують практичну цінність аналізу тональності для платформ електронної комерції — зокрема, для виявлення незадоволених клієнтів, удосконалення товарних пропозицій і персоналізації маркетингових стратегій. Робота робить внесок у розвиток багатомовного аналізу тональності в умовах обмежених мовних ресурсів.

Наукова новизна дослідження полягає в поєднанні багатомовних трансформерів із адаптивними класифікаторами та спеціальними методами боротьби з дисбалансом класів для обробки реальних даних ринку електронної комерції. Запропонований підхід може бути адаптований до інших мов і предметних областей, що відкриває перспективи подальших міждисциплінарних досліджень у сфері цифрового маркетингу, онлайн-комунікацій та соціальних медіа.

Ключові слова: *сентимент аналіз, обробка природної мови, електронна комерція, клієнтський досвід, багатомовні трансформерні моделі, XLM-RoBERTa, дисбаланс класів.*