Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data

Yurii Shynkarov
Ukrainian Catholic University
Lviv, Ukraine
shynkarov.pn@ucu.edu.ua

Veronika Solopova
Technische Universität Berlin
Berlin, Germany
veronika.solopova@tu-berlin.de

Vera Schmitt
Technische Universität Berlin
Berlin, Germany
vera.schmitt@tu-berlin.de

Abstract

This paper addresses the challenges of sentiment analysis in Ukrainian social media, where users frequently engage in code-switching with Russian and other languages. We introduce COSMUS (COde-Switched MUltilingual Sentiment for Ukrainian Social media), a 12,224-texts corpus collected from Telegram channels, product-review sites and open datasets, annotated into positive, negative, neutral and mixed sentiment classes as well as language labels (Ukrainian, Russian, code-switched). We benchmark three modeling paradigms: (i) few-shot prompting of GPT-4o and DeepSeek V2-chat, (ii) multilingual mBERT, and (iii) the Ukrainian-centric UkrRoberta. We also analyze calibration and LIME scores of the latter two solutions to verify its performance on various language labels. To mitigate data sparsity we test two augmentation strategies: back-translation consistently hurts performance, whereas a Large Language Model (LLM) word-substitution scheme yields up to +2.2% accuracy. Our work delivers the first publicly available dataset and comprehensive benchmark for sentiment classification in Ukrainian code-switching media. Results demonstrate that language-specific pre-training combined with targeted augmentation yields the most accurate and trustworthy predictions in this challenging low-resource setting.

Disclaimer: our figures include attested linguistic occurrences of non-normative lexicon.

1 Introduction

Sentiment analysis has long been one of crucial tasks in natural language processing (NLP), with wide-ranging applications in business, media, and the social sciences. The field saw significant progress with the adoption of deep learning techniques and the introduction of transformer-based architectures, which enabled state-of-the-art sentiment classifiers to routinely achieve over 90% accuracy and F1-scores on English-language benchmarks (Mao et al., 2024; ben, 2024). However, these advancements are unevenly distributed, as high-resource languages benefit from abundant labeled data. This gap is especially pronounced in informal, multilingual settings such as Ukrainian social media, where users frequently mix dialects, use transliterations, and code-switch with Russian and other languages. This involves not only mixing lexicon and morphems, but also grammatical forms and structures between several languages within one linguistic utterance or text (Poplack, 1980). With nearly 20% of users engaging in content beyond Ukrainian (Raz, 2024), there is a clear need for sentiment analysis systems that are multilingual and code-switching aware.

To address these gaps, we propose and test a comprehensive framework for sentiment analysis in Ukrainian social media, tailored to the unique linguistic landscape. Our contributions are threefold:

(1) We develop a high-quality, annotated dataset of Ukrainian social media content that includes labels for both sentiment and language. The dataset [1], code [2] and models [3] are accessible under under CC BY 4.0 (Attribution).

(2) We evaluate various augmentation strategies—LLM word-substitution scheme and back-translation—for improving sentiment classification under low-resource constraints.

(3) We fine-tune and benchmark small transformer-based architectures on our dataset, and compare their performance against general-purpose LLMs in zero- and few-shot setups.

In addition, we apply an explainable AI (XAI) LIME analysis (Ribeiro et al., 2016) and model cal-

[1] https://osf.io/2m6et/files/osfstorage

[2] https://github.com/ShynkarovUCU/UASocialSentiment

[3] https : / / huggingface . co / YShynkarov / ukr-roberta-cosmus-sentiment

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 179–193
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

ibration analysis to verify the reliability of our classifier approaches. Our findings contribute to both the Ukrainian NLP landscape and the broader field of sentiment analysis in low-resource and code-mixed language environments.

2 Related Work

Most previous studies on sentiment analysis in code-switching linguistic settings focused on Spanish and English (Sp-Eng CS) (Aryal et al., 2022; Vilares et al., 2016) and the variable linguistic landscape of Indian languages (Ahmad et al., 2022a,b), including their intra-sentential code-switching with English, such as in the case of Dravidian (Prakash and Vijay, 2024). While code-switching in Ukrainian has been increasingly studied across various genres—including parliamentary discourse (Kanishcheva et al., 2023), mixed-speech transcripts (Pylypenko and Lyudovyk, 2019), and social media platforms (Orobchuk, 2024)—few studies directly address the problem of sentiment analysis.

Existing sentiment analysis research in Ukrainian primarily focuses on monolingual contexts or uses Russian as a dominant language. Bobichev et al. (2017) explore sentiment trends in Ukrainian and Russian news articles, applying lexicon-based techniques, while Romanyshyn (2013) present a rule-based method for analyzing user reviews written in Ukrainian. More recent datasets, described e.g. in Baida (2023) and Ustyianovych and Barbosa (2024), incorporate mixed-language content; however, their primary focus remains on Russian-dominant corpora or use sentiment orientations related to political stance rather than emotion polarity.

Entity-level sentiment classification has been applied in Ukrainian-language media (Makogon and Samokhin, 2021), demonstrating the viability of transformer-based models fine-tuned on domain-specific data. Yet these approaches often assume standardized language inputs, omitting the hybrid linguistic characteristics seen on platforms like Telegram, where code-mixing, dialectal variation, and transliteration are common.

While general-purpose multilingual models like mBERT and XLM-R have been applied to sentiment analysis in low-resource European languages (Filip et al., 2024; Vileikytė et al., 2024), their robustness in Ukrainian-Russian code-switched settings remains unexplored. Recent experiments began to explore fine-tuning large multilingual transformers or LLMs (e.g., GPT-4, LLaMA3) for this task (Buscemi and Proverbio, 2024; Ustyianovych and Barbosa, 2024), with mixed results and limited evidence of generalization to informal social media discourse. In summary, although foundational work exists for sentiment detection in Ukrainian, there remains a notable absence of approaches tailored to the challenges of code-switching.

3 Methodology

To address the identified gaps, we propose a sentiment classification approach for Ukrainian social media data, encompassing data preprocessing, annotation, and a structured experimental methodology. In this study, we do not differentiate between code-switching (intersentential) and code-mixing (intra-sentential) and refer to the phenomenon as a whole as code-switching.

3.1 Data Preprocessing

We constructed our dataset partially from publicly available datasets, namely TG from Baida (2023) with 3,000 samples and 1,000 Yakaboo book reviews[4]. Additionally, we scraped posts and comments on Ukrainian social media channels from Telegram, collected between February 2022 and September 2024 (8,064) and product reviews from Hotline.ua (1,000 texts). After deleting duplicates and overly short utterance, the initial corpus resulted in 12,224 documents spanning diverse topics such as politics, governmental services, entertainment, daily life, and online reviews of books and marketplaces. The average length of a text in the dataset is 170 characters, while the median length is 96 characters. The dataset also contains 7% of longer texts exceeding 500 characters. 28% of texts contain emojis reflecting the colloquial nature of the corpus. The data was anonymised to exclude personal information. All personal and sensitive data were removed from the texts, such as banking card numbers, addresses, personal emails, full names and web links using regex matching.

To ensure representation of code-switching phenomena, we employed GPT-4o (OpenAI, 2024) model using OpenAI API and *lang-detection*(Shuyo, 2010) to detect if a text is monolingual (Ukrainian or Russian, other) or code-switched. If the language-detector predicted Ukrainian, we chose Ukrainian as a label, because

[4] https://github.com/osyvokon/awesome-ukrainian-nlp

Label	Precision (GPT)	Recall (GPT)	Precision (Hybrid)	Recall (Hybrid)	Count
Ukrainian	0.967	0.696	0.974	0.904	125
Russian	0.909	0.690	0.824	0.966	58
Code-mixed	0.197	0.765	0.812	0.765	17

Table 1: Precision and recall per language label (n=200). Hybrid means GPT & language-detection results.

this detector was shown to have high precision for this language. If it predicted other languages, we chose the GPT label. During this process, we filtered out all texts in languages other than Ukrainian and Russian (primarily English and Polish) because their presence in the dataset was statistically insignificant and would not contribute meaningfully to our analysis of code-switching patterns. The resulting dataset includes monolingual Ukrainian, monolingual Russian, and code-switched content in proportion of 66%, 28% and 6% respectively.

We manually validated a subset of 200 samples of automatic language annotations, randomly chosen to to represent same language proportions as in the full dataset. The results of the co-annotation can be seen in Table 1 for both pure GPT and hybrid GPT and language-detection results. Overall, in the case of the GPT model, it identifies mixed well when it is truly present (high recall), but it over-predicts it the cases of miss-spellings (low precision), while Ukrainian and Russian, are moderately well-predicted. However, with our hybrid approach we achieved high results for all of the language settings, and especially improved code-mixed results.

3.2 Data Annotation

To facilitate the annotation process, we developed a dedicated Telegram bot to distribute annotation guidelines and collect annotators' responses. Five annotators, all native Ukrainian speakers with bilingual proficiency in Russian participated. The annotation guidelines instructed annotators to classify texts according to four sentiment categories: positive, negative, neutral and mixed sentiment. The guidelines emphasized that sentiment classification should be based on specific expressions present in the text rather than the annotator's subjective interpretation of the author's intent. We provided multiple examples to illustrate each category, including edge cases where the factual content might seem negative, but the text itself contains no sentiment-bearing expressions and should be classified as neutral. The annotation guidelines can be found in Appendix B in original Ukrainian version and English translation. Annotators were also instructed to identify spam messages and mark them for deletion from the dataset. We used "I do not know" label for such cases, and filtered these data points in post-processing. This additional filtering step helped ensure the quality and relevance of our final corpus.

To establish consistency and measure inter-annotator agreement, we designed the annotation process so that the first 100 texts were identical for all five annotators. This overlap allowed us to calculate Cohen's kappa for sentiment labels. The average result for all annotators is ($\kappa = 0.79$), indicating substantial agreement. Disagreements were resolved with majority voting during the final preprocessing steps. The final sentiment distribution in the annotated dataset can be found in Table 2.

Sentiment	Count	Percentage
Neutral	4,702	38%
Negative	4,541	37%
Positive	2,373	19%
Mixed	608	6%
Total	**12,224**	**100%**

Table 2: Sentiment distribution of the dataset.

Finally, we divided the dataset into training (80%) and test (20%) sets, maintaining the distribution of sentiment and language categories across splits while also controlling for text length distribution to account for the observed skewness towards longer texts.

3.3 Experimental Setup

In this section, we describe the established LLM baseline and the fine-tuning process.

Prompting Strategy. We implemented GPT-4o (OpenAI, 2024) and Deepseek V2-chat (DeepSeek-AI, 2024) as our prompting-based baselines, conducting several experiments to maximize performance. The general approach was to structure the prompt to include the same sentiment definitions, edge cases, and decision criteria used by human annotators. We tested writing prompts in both Ukrainian and English (see final prompt in Appendix A).

Fine-tuning Approach. As sentiment analysis has multiple benefits for business analytics, we also fine-tuned two transformer-based Small Language Models (SLMs) from the BERT family as more cost-effective deployment solutions:

(1) **UkrRoberta** (Radchenko, 2021): A model additionally pre-trained on Ukrainian text data with Roberta architecture, optimized for Ukrainian language understanding.

(2) **Modern BERT (mBERT)** (Warner et al., 2024): A multilingual BERT variant optimized for cross-lingual transfer across various languages, including Ukrainian and Russian.

For each SLM model, we implemented a classification head on top of the pre-trained transformer architecture. To handle longer texts that exceeded the maximum token length, we employed a segmentation approach where texts were divided into sections matching the maximum token length. Predictions were made for each segment, and the final classification was determined through majority voting across segments. We utilized Optuna (Akiba et al., 2019) for systematic hyperparameter tuning.

Data Augmentation. To address potential data sparsity, particularly for code-switched content, we experimented with two augmentation strategies:

(1) **Back-translation**: translating[5] text to an intermediate language (English) and back to the original language (Ukrainian or Russian) to generate paraphrased alternatives while preserving sentiment.

(2) **Word substitution**: using gpt-4o we replaced words with synonyms or contextually appropriate alternatives while maintaining the original sentiment and code-switching patterns.

For the second strategy, we employed the GPT-4o model to perform word substitutions, with a particular emphasis on preserving sentiment. The model was accessed via API with a temperature setting of 0.7 to produce diverse yet contextually appropriate replacements. We used in-context learning, providing explicit examples of the desired substitution patterns within the prompt. The model was instructed to recognize and preserve code-switching patterns while making lexical substitutions, and to maintain the original sentence structure (see Appendix D). We performed a sentiment consistency check by manually reviewing a statistically significant subset of newly generated samples from each sentiment class.

The augmentation ratio was class-dependent, with higher ratios for minority classes and lower ratios for well-represented classes. The overall goal was to improve the class balance in the original dataset.

Evaluation Methodology. We use standard metrics such as precision, recall, F1-score (micro & macro) and accuracy, to evaluate the classification task while accounting for class imbalance in the created dataset. We also measure Expected Calibration Error (ECE) from Nixon et al. (2019) to assess the reliability of the SLM solutions, specifically applied to different language subsets, computed as:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\mathrm{acc}(b) - \mathrm{conf}(b)|, \qquad (1)$$

where B is the number of bins, n_b is the number of predictions in bin b, and N is the total number of data points. Each prediction is assigned to a bin based on its confidence score (i.e., the predicted probability of the top class), and acc(b) and conf(b) denote the average accuracy and average confidence within bin b, respectively.

4 Results

4.1 Data Augmentation Results

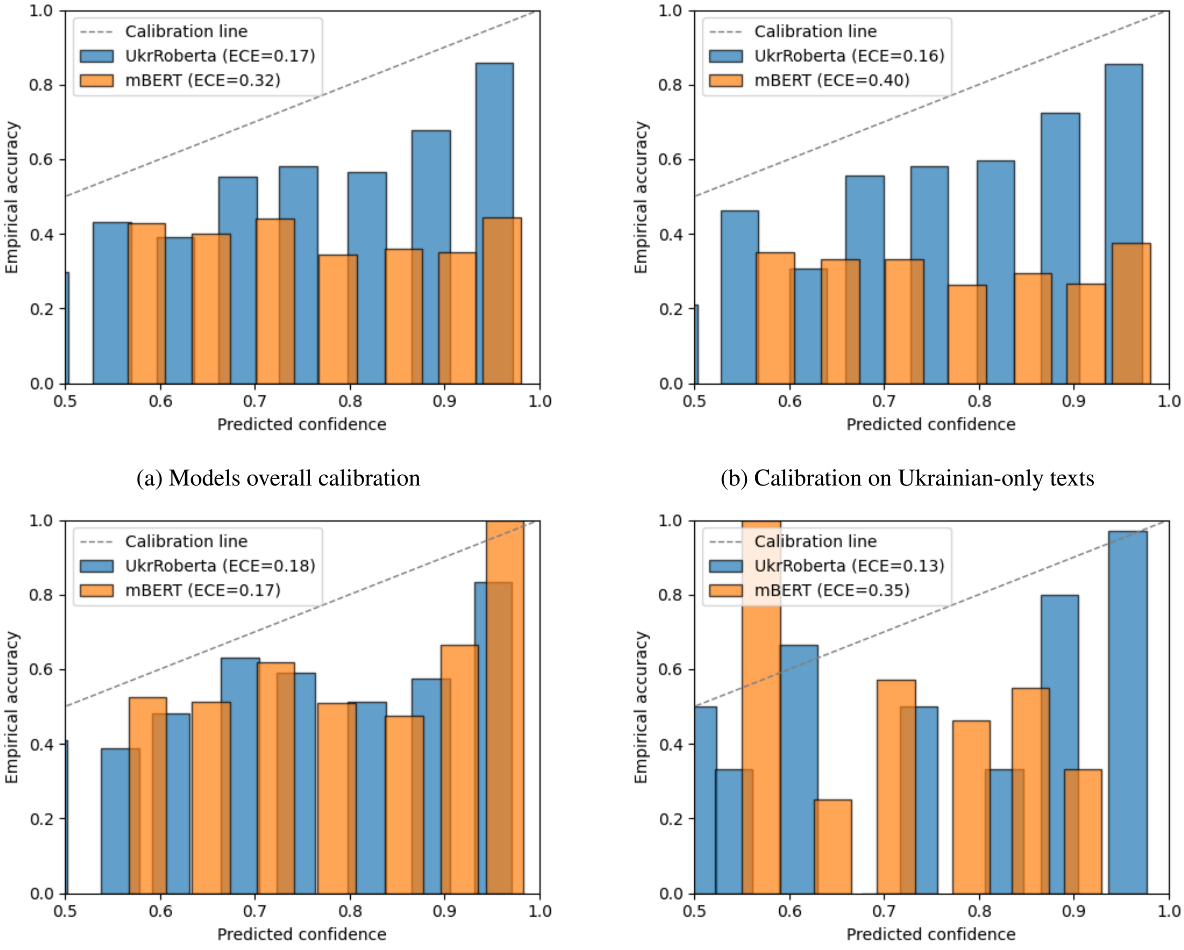
We evaluated our proposed approach using three configurations. Table 4 presents the accuracy results for GPT-4o, DeepSeek V2-chat, mBERT, and UkrRoberta on the original dataset and two augmented datasets.

The effects of the data-augmentation strategies varied across models. The word-substitution strategy, which preserves code-switching patterns and text structure while introducing lexical variety, proved to be a valuable training signal for SLM models. Back-translation, however, consistently degraded performance for all models, with decreases of 3.2% for GPT-4o, 2.7% for DeepSeek, 2.3% for mBERT, and 2.2% for UkrRoBERTa. This degradation likely stems from the loss of contextual cues and code-switching patterns during the translation process.

[5] For translations, we used LibreTranslate, an open-source neural machine translation tool (Klein et al., 2017).

		UkrRoberta			mBERT		
Language	**Metric**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
UA	Macro	0.67	0.61	0.63	0.73	0.44	0.43
	Micro	0.74	0.74	0.73	0.64	0.57	0.54
RU	Macro	0.58	0.60	0.59	0.81	0.61	0.66
	Micro	0.71	0.71	0.71	0.77	0.74	0.74
Code-Switched	Macro	0.72	0.69	0.68	0.69	0.51	0.54
	Micro	0.76	0.69	0.71	0.80	0.58	0.60
Overall	Macro	0.66	0.62	0.64	0.80	0.58	0.58
	Micro	0.74	0.74	0.73	0.73	0.69	0.67

Table 3: Performance comparison between UkrRoberta and mBERT sentiment classification models. Word-substitution augmentation is applied for both models. Macro metrics calculate the unweighted average across classes, while micro metrics account for class imbalance.



(a) Models overall calibration

(b) Calibration on Ukrainian-only texts

(c) Calibration on Russian-only texts

(d) Calibration on code-switched texts

Figure 1: Reliability diagrams for UkrRoberta and mBERT calibration across language subsets

4.2 Overall Performance

UkrRoberta demonstrated the strongest overall performance, achieving 73.6% accuracy with word substitution augmentation, a significant improvement over both the few-shot prompting approach (70.3% for GPT-4o, the multilingual mBERT (69.8%), and deepseek, which showed the lowest results (64.6%). This finding suggests that language-specific pre-training offers substantial benefits for sentiment analysis in Ukrainian social-media contexts, particularly when handling code-switched content.

Model	Original	Back-transl.	Word subs.
GPT-4o	**70.3%**	67.1%	68.2%
DeepSeek V2-chat	**64.6%**	61.9%	62.7%
mBERT	68.8%	66.5%	**69.8%**
UkrRoberta	71.4%	69.2%	**73.6%**

Table 4: Accuracy (%) of sentiment classification models across different data augmentation strategies. Best results per model are in bold.

4.3 Performance Across Language Categories

To assess the robustness of our SLM models across language categories, we evaluated their performance separately on Ukrainian monolingual, Russian monolingual, and code-switched texts.

As shown in Table 3, we observe distinct performance patterns across language categories. For Ukrainian monolingual and code-switched texts, UkrRoberta outperforms mBERT, posting higher micro F1-scores — 0.73 vs 0.54 and 0.71 vs 0.60, respectively. The pattern reverses for Russian texts, where mBERT is stronger (0.74 vs 0.71 micro F1). Notably, mBERT also achieves relatively high precision on Russian content (0.81 macro).

A clear precision–recall trade-off emerges. While mBERT generally delivers higher precision, UkrRoberta offers a more balanced precision–recall profile and superior recall. This balance is valuable for applications in the domain under study, where false negatives and false positives incur comparable costs.

Models Calibration. In addition, we assessed the reliability of the SLMs' sentiment predictions by computing the ECE for each model and each language. We then plotted the corresponding reliability diagrams to show how closely the models' confidence scores track the true likelihood of correctness (see Figure 1).

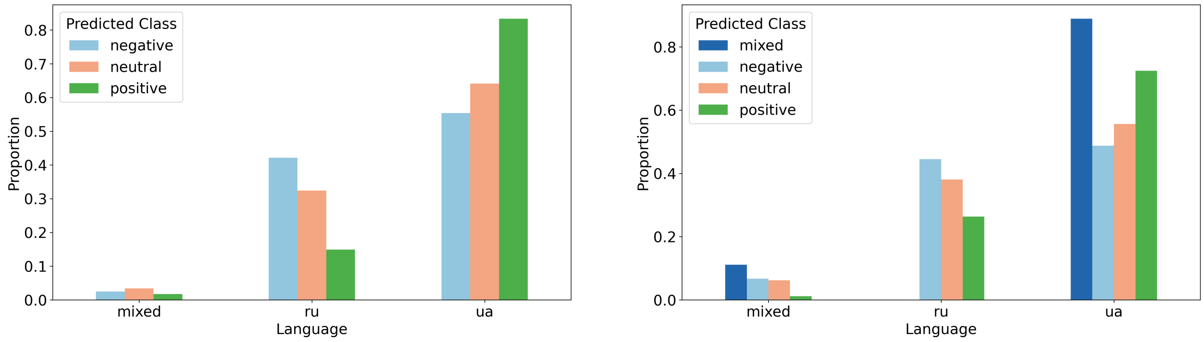
Across the full test set, UkrRoberta exhibits substantially better calibration (ECE = 0.17) than mBERT (ECE = 0.32), with bars that track the Calibration line more closely in every bin. A similar pattern emerges for monolingual Ukrainian (ECE = 0.16 vs 0.40) and code-mixed texts (0.13 vs 0.35), additionally underscoring the benefit of language-specific pre-training. The trend is only slightly reversed for Russian-only inputs: mBERT's ECE of 0.17 marginally surpasses UkrRoberta's 0.18, mirroring mBERT's higher precision on this subset.

4.4 Explainability

Another facet of our research was identifying the sentiment-bearing linguistic features captured by the best-performing classifier. We calculated LIME scores for the test-set texts under the two best UkrRoberta configurations, as the best performing model overall: three-class and four-class. We then examined the tokens with the highest LIME scores for each language and class. By comparing correct and incorrect classifications, we also analyzed the tokens that most frequently caused confusion (see Figure 3). Finally, we verified potential language bias by measuring how often tokens from each language category — Ukrainian, Russian, and Code-Switched — contributed positively to each class prediction.

Language bias. As it is illustrated in Figure 2, both best settings of UkrRoberta exhibit certain language bias against Russian tokens, more often attributing them strong negative bias, while Ukrainian tokens are more prone to contribute to positive, or mixed sentiment predictions, in case of the 4-class model. Code-switched subsets' tokens contribute more often to mixed sentiment predictions, but otherwise show rather well-distributed terms over neutral and negative classes, but are the least prone to contribute to positive predictions. However, it is inherently more complicated to analyse tokens from code-switched subset, as they can include both code-mixed and standard Ukrainian or Russian tokens.

Term importance. As for the highest-scoring terms according to the LIME analysis, the 3-class and 4-class UkrRoberta show overall similar patterns. Top terms biasing predictions toward the **negative class** (Figure 3 (a) and (b)) include non-normative lexicon, war-related vocabulary, such as ukr. "розбомбленная" (en. bombed-out), ru. "хуячит" (profanity for shelling), and ru. "обстреливают" (en. they are shelling); terms associated with Russian or non-democratic identity, such as ukr. "вата" (en. "cotton"— derogatory slang for pro-Russian individuals), "русня" (derogatory term for Russians), "відкат" (en. rollback, reversal, kickback used in relation to reforms and positive social changes), "підозра" (suspici-on); and adjectives with negative connotation, such as ukr. "жахливий" (en. horrible), "гнилий" (en. rotten), and "холодний" (en. cold). Interestingly, both models assign high importance to onomatopoeic laughter tokens, suggesting that the mod-



(a) 3-class model LIME analysis

(b) 4-class model LIME analysis

Figure 2: Language contribution of the test set to predicted sentiment classes with LIME score ≥ 0.1. The left plot shows results for a 3-class setup (negative, neutral, positive), and the right plot shows a 4-class setup including mixed. The results are normalised by the sapmple size to minimise influence of the varying class and language representation in the test set.

els are able to detect irony. However, our LIME analysis for incorrect predictions of the Ukrainian positive class conflicted with this findings, since such laughter tokens actually contributed to misclassifying instances as positive (see Figure 3 (h)). Overall, the evidence for confusing terms in this class is inconclusive. Both models may suffer from a common issue in sentiment analysis, where the presence of negations is overly attributed to negative polarity. In the 4-class model, we also observed terms that typically have a clearly negative connotation being flagged as confusing, indicating they may have been used in ironic contexts.

The **neutral class** (Figure 3 (c) and (d)) displays a wider range of confusing terms for the models. This can be attributed to the nature of the words themselves — such as conjunctions and emotionally neutral verbs and nouns — which may lead the model to classify inputs as neutral based on the absence of emotionally charged terms rather than the presence of neutral ones.

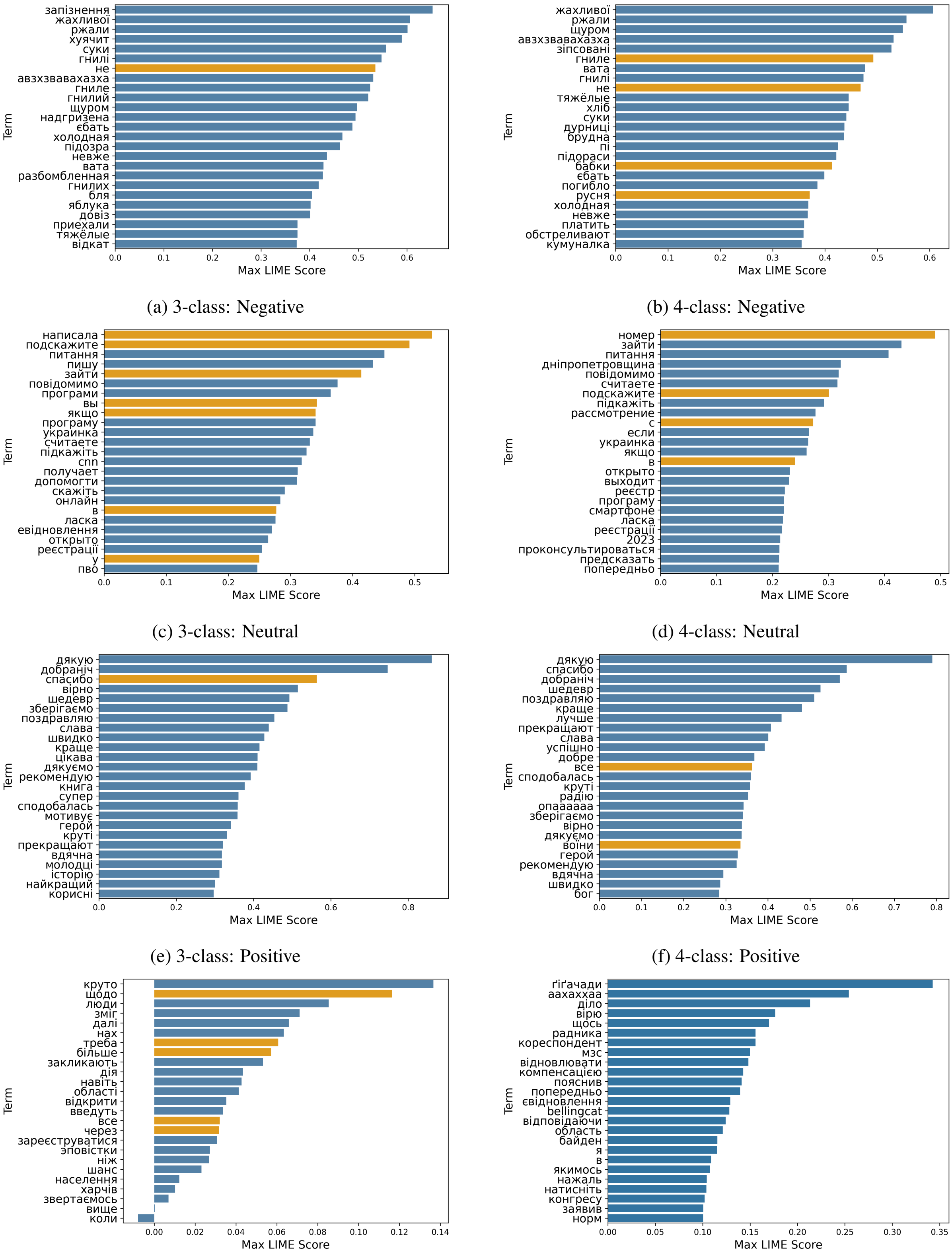
Terms contributing to the **positive class** prediction (Figure 3 (e) and (f)) show fewer confusing cases. In the 3-class model, this may again reflect ironic expressions of gratitude. In the 4-class model, however, we observed a fatalistic use of ukr. "все" (en. everything, that's all / enough) and a mixture of tragic and heroic contexts containing ukr. "воїни"(en. soldiers), which might contribute to the model's uncertainty. Specifically for Ukrainian, we also observed that many conventionally positive words used in ironic or sarcastic colloquial contexts are not well captured by the model, such as ukr. "ґіґачади" (en. giga-chads), "діло" (matter), "вірю" (en. I believe). Additionally, many terms connected to governmental institutions or proper nouns like "Biden" or "Bellingcat" are among confusing, reflecting the pluralism of political opinions expressed in the training data.

Finally, the **mixed sentiment class** of the 4-class model, illustrated in Figure 3 (g), shows a predominantly neutral lexicon. The most notable exceptions are a strongly negative expressive profanity marker ru. "нах" (shortened vulgar form of go to hell) and a colloquial positive qualifier ukr. "круто" (en. cool, awesome). However, there is insufficient evidence to claim that the model has learned the concept of mixed sentiment from the data.

5 Discussion

Overall, UkrRoberta's stronger performance (0.73 vs 0.67 micro F1) confirms that language-specific pre-training, combined with targeted word-substitution augmentation, is a more effective strategy for sentiment analysis in the linguistically complex landscape of Ukrainian social media. While our peak accuracy of 73.6% is lower than the 90%+ performance often reported for monolingual English sentiment analysis systems (Mao et al., 2024; ben, 2024), the performance relationship we observe between general-purpose LLMs and smaller, task-specific fine-tuned models aligns with findings from prior work(Barbieri et al., 2022; Filip et al., 2024). This indicates that our approach performs comparably to existing solutions despite the inherent complexities of Ukrainian-Russian code-switching in social media content.

Our calibration analysis findings confirm that good discrimination does not automatically entail



(a) 3-class: Negative

(b) 4-class: Negative

(c) 3-class: Neutral

(d) 4-class: Neutral

(e) 3-class: Positive

(f) 4-class: Positive

(g) 4-class: Mixed

(h) 4-class: UA ’positive’ (wrong predictions)

Figure 3: Top LIME terms contributing to sentiment predictions across classes and model variants. Each row compares the same class across the 3-class and 4-class models: (a,b) Negative, (c,d) Neutral, (e,f) Positive. Row 4 includes (g) the Mixed class (only in 4-class) and (h) the top terms associated with misclassified Ukrainian-language examples predicted as ’positive’. Orange bars indicate terms shared with incorrect predictions, potentially contributing to false positives or false negatives. Terms are case-normalized; for repeating terms, only the highest LIME score is retained.

good calibration: while mBERT achieves competitive F1 on Russian texts, its reliability sharply degrades on Ukrainian and mixed inputs. Conversely, UkrRoberta delivers more trustworthy probability estimates in the linguistically diverse conditions typical of Ukrainian social media. While bias is generally undesirable, some degree of bias may be contextually appropriate in Ukrainian historical and political context. Although UkrRoBERTa slightly underperforms in Russian and exhibits a tendency to associate Russian lexical items with negative sentiment, this trade-off seems more acceptable than reversed mBert's scenario, more favorable towards the Russian language. Since AI naturally amplifies and re-enforces existing biases, and considering that Ukrainian language is historically downplayed and discriminated against, the choice between mBert and UkrRoberta should involve these additional socio-linguistic considerations. Additionally, UkrRoberta may reflect real-world patterns of usage and sociopolitical framing of sentiment in Ukrainian wartime discourse. Finally, we storngly advocate that interpretability and calibration are essential in evaluating sentiment models beyond F1 scores—especially when language identity and political stance are intertwined. While our best-performing model (UkrRoBERTa with word substitution) shows promising robustness, further work is needed to handle sarcasm, negation, and mixed affect more reliably.

6 Conclusion

We present COSMUS, the first publicly available, 12,224 texts corpus of Ukrainian, Russian and code-switched social media texts with four-way sentiment labels and substantial annotator agreement. Fine-tuning the UkrRoBERTa with GPT-4o–driven data augmentations yields the top accuracy of 73.6%, surpassing mBERT and few-shot LLM baselines. Reliability diagrams and LIME analysis show UkrRoBERTa is also better calibrated across most language subsets and exhibits less language bias on Ukrainian and code-mixed samples.

Limitations

While this study contributes a novel dataset and modeling pipeline for sentiment analysis in Ukrainian code-switching contexts, several limitations must be acknowledged. Despite our efforts to include diverse sources and augment underrepresented classes, code-switched texts still constitute only 6% of the COSMUS dataset, which does not perfectly reflect Ukrainian social media reality and limits the robustness of model generalization on code-switching phenomena. The manual validation results indicate that the real number of code-switched samples may be even lower (low precision). This imbalance may limit the model's ability to generalize to real-world social media contexts, where hybrid and fluid language use is more prevalent. Future data collection efforts should aim for more representative sampling of code-switched communication. Moreover, the exclusion of other relevant language pairs (e.g., Ukrainian–English or Ukrainian–Polish) restricts the broader applicability of our findings to multilingual contexts beyond Russian–Ukrainian.

Although we ensured substantial inter-annotator agreement (κ = 0.79), the classification of subtle or sarcastic sentiment—especially in politically charged or ironic discourse—remains subjective. While the use of concrete sentiment-bearing expressions mitigates this, future work could benefit from multi-layered annotation schemes or continuous sentiment scales. Bigger data overlap between annotators would also be beneficial.

Even our best-performing model, UkrRoberta with word substitution, struggles with sarcasm, negation, and mixed emotions, as evidenced by LIME analyses and misclassifications. This reflects broader challenges in sentiment modeling across informal, affectively ambiguous genres. The detected language bias, wherein Russian tokens are more frequently associated with negative sentiment, raises important ethical and interpretability questions. While we contextualize this as potentially reflecting real-world sociopolitical dynamics, further research is needed to disentangle model-internal bias from corpus-driven patterns, especially when deploying such models in sensitive applications.

Finally, while this study primarily focused on platforms with a pro-Ukrainian or neutral stance, many globally influential information ecosystems include actors and communities with hostile or adversarial messaging toward Ukraine. Excluding these from the current analysis may limit the broader validity of our findings. Future research should expand the scope of sentiment modeling to include content from such platforms to better understand and model the full spectrum of narratives shaping public discourse in and about Ukraine.

References

2024. Nlp-progress: repository to track the progress in natural language processing (nlp), including the datasets and the current sota. Accessed: 2025-01-26.

2024. Sociological survey: identity of ukrainian citizens and trends of change. Accessed: 2025-01-26.

Gazi Ahmad, Jimmy Singla, Anis Ali, Aijaz Reshi, and Anas A. Salameh. 2022a. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus - a comprehensive review. *International Journal of Advanced Computer Science and Applications*, 13.

Gazi Ahmad, Jimmy Singla, Anis Ali, Aijaz Reshi, and Anas A. Salameh. 2022b. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus - a comprehensive review. *International Journal of Advanced Computer Science and Applications*, 13.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Saurav K. Aryal, Howard Prioleau, and Gloria J. Washington. 2022. Sentiment classification of code-switched text using pre-trained multilingual embeddings and segmentation. *ArXiv*, abs/2210.16461.

Dmytro Baida. 2023. Autotrain dataset for project: ukrainian-telegram-sentiment-analysis. Accessed: 2024-01-26.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Victoria Bobichev, Olga Kanishcheva, and Olga Cherednichenko. 2017. Sentiment analysis in the ukrainian and russian news. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pages 1050–1055.

Alessio Buscemi and Daniele Proverbio. 2024. Chatgpt vs gemini vs llama on multilingual sentiment analysis. *Preprint*, arXiv:2402.01715.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Tomáš Filip, Martin Pavlíček, and Petr Sosík. 2024. Fine-tuning multilingual language models in twitter/x sentiment analysis: a study on eastern-european v4 languages. *Preprint*, arXiv:2408.02044.

Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova, and Ruprecht von Waldenfels. 2023. The parliamentary code-switching corpus: Bilingualism in the Ukrainian parliament in the 1990s-2020s. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 79–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Iuliia Makogon and Igor Samokhin. 2021. Targeted sentiment analysis for ukrainian and russian news articles. In *ICTERI-2021, Vol II: Workshops*, pages September 28 – October 2, Kherson, Ukraine. CEUR-WS.org, CEUR Workshop Proceedings.

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4):102048.

Jeremy Nixon, Mike Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. *CoRR*, abs/1904.01685.

OpenAI. 2024. Gpt-4o technical report. Accessed: 2025-04-17.

Dariia Orobchuk. 2024. Charting language shift through ukraine's social media actors. *Canadian Slavonic Papers*, 66(3-4):431–455.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en español: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

V. Jothi Prakash and S. Arul Antran Vijay. 2024. A novel socio-pragmatic framework for sentiment analysis in dravidian–english code-switched texts. *Knowledge-Based Systems*, 300:112248.

Valeriy Pylypenko and Tetyana Lyudovyk. 2019. Automatic recognition of mixed ukrainian-russian speech. In *Proceedings of the International Conference on Language Technologies for All (LT4All)*. UNESCO.

Vitalii Radchenko. 2021. youscan/ukr-roberta-base. `https://huggingface.co/youscan/ukr-roberta-base`. Accessed: 2025-04-17.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Mariana Romanyshyn. 2013. Rule-based sentiment analysis of ukrainian reviews. *International Journal of Artificial Intelligence & Applications*, 4(4).

Nakatani Shuyo. 2010. Language detection library for java.

Taras Ustyianovych and Denilson Barbosa. 2024. Instant messaging platforms news multi-task classification for stance, sentiment, and discrimination detection. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 30–40, Torino, Italia. ELRA and ICCL.

David Vilares, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2016. EN-ES-CS: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153, Portorož, Slovenia. European Language Resources Association (ELRA).

Brigita Vileikytė, Mantas Lukoševičius, and Lukas Stankevičius. 2024. Sentiment analysis of lithuanian online reviews using large language models. *Preprint*, arXiv:2407.19914.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

A Baseline Solution Best Performing Prompt

You are an expert in determining the sentiment of a text. Our task is to determine the emotion that a person puts into a written text as accurately as possible. To do this, I will show you texts from Ukrainian social networks, and you will choose the correct answer regarding the sentiment. The answer options will be as follows:

1. Positive -> expressions used that reflect positive emotions (joy, support, admiration, etc.);

2. Negative -> expressions used that reflect negative emotions (criticism, sarcasm, condemnation, aggression, doubt, fear, etc.);

3. Neutral -> the author does not use either positive or negative expressions (neutral emotion);

4. Mixed -> the text contains expressions from both the positive and negative spectrum of emotions (mixed case);

It is important that you do not indicate your own guess about the author's sentiment, but find indications of it in specific expressions. I will give a few examples.

Examples:

"Аварії > this short text has a neutral sentiment. Despite the fact that the Ukrainian word "Аварії "often has a negative context, in this case there is no additional information reflecting the sentiment of the author.

"Так я ж тебе задал вопрос. Киев, май, первое применение пэтриотов - когда все небо осветили этим - были там и х22, и кинжалы - так были прилеты тогда? Не было. Вопрос залу - почему так произошло? Пэтриоты сбивают всю эту срань > this text has a negative sentiment. The author uses expressions that characterize aggression and criticism of the interlocutor.

"Зникло світло у Святошинському районі. > this text has a neutral sentiment. The fact of the lack of electricity itself is perceived negatively, but the author of the text does not use either positive or negative words / expressions.

"Проблеми зі світлом в Києві та області після вибухів! > in turn, the following news item has a negative connotation. The author demonstrates his attitude through the word "Проблеми"and the exclamation mark "!", emphasizing the expression.

":cry: Внаслідок ракетної атаки зафіксовано падіння уламків в Печерському районі на дах багатоповерхового житлового будинку, – КМВА > text with a negative sentiment, which the author demonstrates through the use of the ":cry:" emoji.

"Ну норм > this is an example of a positive sentiment. The text itself is not very expressive, but the author clearly demonstrates the emotion of "approval" of something, which belongs to the positive spectrum.

":exclamation:В бік Києва пуски ще декількох 'Кинджалів'. Ворог намагається пробити наші ППО. Поки відбиваємося, але є падіння уламків, тож перебуваємо в укриттях або хоча б за парою стін.> this news item is an example of a negative sentiment. The author demonstrates his attitude to the event through the expressions "Ворог намагається пробити наші ППО. Поки відбиваємося, але ",

Your answer should be only one word. THIS IS IMPORTANT! You must answer exclusively with only one word from the list: [positive, negative, neutral, mixed].

B Annotation Guidelines

B.1 Original Ukrainian Guidelines

Наше завдання - навчитись визначати емоцію (сентимент), яку людина закладає у написаний текст. Для цього бот показуватиме тобі тексти з українських соціальних мереж, а ти - обиратимеш вірний варіант відповіді щодо сентименту. Варіанти відповідей будуть наступні:

1. Позитивний -> Використані вирази, що відображають позитивні емоції (радість, підтримку, захоплення тощо);

2. Негативний -> Використані вирази, що відображають негативні емоції (критика, сарказм, осуд, агресія, сумнів, страх тощо);

3. Нейтральний -> Автор не використовує ні позитивних, ні негативних виразів (нейтральна емоція тексту);

4. Змішаний -> Текст містить вирази як з позитивного спектру емоцій, так і з негативного (змішаний випадок);

5. Я не впевнений -> Дану опцію слід обрати, якщо ти не впевнений у правильності вибору.

Важливо, що потрібно вказувати не власну здогадку щодо сентименту автора, а знаходити вказівки на нього у конкретних виразах. У наступному пості надам декілька прикладів Приклади:

1. "Аварії > цей короткий текст має нейтральний сентимент. Попри те, що слово "аварії" часто має негативний контекст, у даному випадку відсутня будь-яка додаткова інформація, що відображає сентимент автора.

2. "Так я ж тебе задал вопрос. Киев, май, первое применение пэтриотов - когда все небо осветили этим - были там и х22, и кинжалы - так были прилеты тогда? Не было. Вопрос залу - почему так произошло? Пэтриоты сбивают всю эту срань > цей текст має негативний сентимент. Попри те, що факт "Петріоти збивають ракети" може відчуватись позитивно, автор використовує вирази, що характеризують агресію та критику до співрозмовника.

3. "Зникло світло у Святошинському районі. > даний текст має нейтральний сентимент. Сам факт відсутності електроенергії сприймається негативно, але автор тексту не використовує ні позитивних, ні негативних слів / виразів.

4. "Зникло світло у Святошинському районі. > даний текст має нейтральний сентимент. Сам факт відсутності електроенергії сприймається негативно, але автор тексту не використовує ні позитивних, ні негативних слів / виразів.

5. "Проблеми зі світлом в Києві та області після вибухів! > у свою чергу наступна новина має негативне забарвлення. Автор демонструє своє відношення через слово "Проблеми" та знак оклику "!", підкреслюючи експресію.

6. "sad emodji Внаслідок ракетної атаки зафіксовано падіння уламків в Печерському районі на дах багатоповерхового житлового будинку, – КМВА > текст із негативним сентиментом, що автор демонструє через використання "sad emodji" емодзі.

7. "Ну норм > це приклад позитивного сентименту. Сам текст не є сильно експресивним, але автор явно демонструє емоцію "схвалення" чогось, яка належить до позитивного спектру.

8. "В бік Києва пуски ще декількох 'Кинджалів'. Ворог намагається пробити наші ППО. Поки відбиваємося, але є падіння уламків, тож перебуваємо в укриттях або хоча б за парою стін. > дана новина є прикладом негативного сентименту. Автор демонструє своє відношення до події через вирази "Ворог намагається пробити наші ППО", "Поки відбиваємося, але...".

9. "С чего ты взял? У меня в Ирпене все окна повыбивало я сохранил квитанцию

то что сам поставил и вернули 20.000 > приклад "змішаного" сентименту. У першій частині автор демонструє критику по відношенню до іншої людини. У другій частині тексту - автор радіє, що йому компенсовано витрати на відновлення домівки.

B.2 English version of the Guidelines

Our task is to learn how to identify the emotion (sentiment) a person conveys in a written text. To do this, the bot will show you posts from Ukrainian social media, and you will choose the correct sentiment classification. The answer options will be as follows:

1. Positive → The text contains expressions that reflect positive emotions (joy, support, admiration, etc.);
2. Negative → The text contains expressions that reflect negative emotions (criticism, sarcasm, condemnation, aggression, doubt, fear, etc.);
3. Neutral → The author does not use either positive or negative expressions (emotionally neutral text);
4. Mixed → The text contains expressions from both the positive and negative emotional spectrum (a mixed case);
5. I'm not sure → Choose this option if you are unsure about the correct sentiment.

Importantly, you should not rely on your guess about the author's sentiment, but instead look for concrete expressions that indicate it. In the next post, I will provide a few examples.

Examples:

1. "Accidents" → This short text has a neutral sentiment. Although the word "accidents" often carries a negative connotation, there is no additional information here that reveals the author's sentiment.
2. "So I asked you a question. Kyiv, May, the first use of Patriots — when the whole sky lit up — there were X-22s and Kinzhals — so were there any hits then? No. Question to the audience — why did that happen? Patriots shoot down all this crap" → This text has a negative sentiment. Although the fact that "Patriots shoot down missiles" might seem positive, the author uses expressions that convey aggression and criticism toward the interlocutor.
3. "Power went out in the Sviatoshynskyi district." → This text has a neutral sentiment. While the fact of a power outage may be perceived negatively, the author uses no clearly positive or negative words or expressions.
4. "Power went out in the Sviatoshynskyi district." → Again, this is a neutral sentiment. Although the situation is unfortunate, the language is emotionally neutral.
5. "Problems with electricity in Kyiv and the region after explosions!" → This post, in contrast, conveys negative sentiment. The word "problems" and the exclamation mark "!" indicate the author's emotional reaction.
6. " sad emodji As a result of a missile strike, debris fell in the Pecherskyi district on the roof of a multi-story residential building, – KMVA" → This is a text with negative sentiment, shown through the use of the "sad emoji" (sad emodji).
7. "Well, okay" → This is an example of positive sentiment. While the expression is not highly emotional, the author clearly shows approval, which falls within the positive spectrum.
8. "Several more 'Kinzhals' launched toward Kyiv. The enemy is trying to break through our air defense. We're still holding them off, but debris is falling, so stay in shelters or behind at least two walls." → This is an example of negative sentiment. The author shows their stance through expressions like "the enemy is trying to break through our air defense" and "we're still holding them off, but...".
9. "Why do you think that? In Irpin, all my windows were blown out — I kept the receipt, did the repairs myself, and got 20,000 back." → This is an example of mixed sentiment. In the first part, the author expresses criticism toward someone. In the second part, the author shows happiness about being reimbursed for repairing their home.

C Prompt For The Word Substitution Augmentation Strategy

You are a sentiment analysis expert. You need to help to create a dataset of texts needed for training an ML model. Your help is to write a text which will be included to the dataset. This is important that the text must language. The sentiment of the text should express sentiment. The example of such a text is provided below.

Write the text similar to the provided example. You MUST do just a rewording. However, remember, that the resulted text must language.

Also, you must write only the text without any additional comments from yourself.

The text example is below: text

D Examples of Manual Language Verification Results

Document Content	GPT	Hybrid	Human
Ну так заметить надо, что получает Краматорск, Дружковка, Славянск, но не Бахмут!(	ru	ru	ru
В Дие есть пункт Євідновлення , там написано что делать	mixed	mixed	mixed
Кастрюлю снять и громко думать що делать((((	mixed	mixed	mixed
У Солом’янському районі уламки ракети впучили у верхні поверхи багатоповерхівки - міський голова	ua	ua	ua
Емм 200 к це якщо квартира пошкоджена чи на будь що. Бо це десь 10% від будинку	mixed	ua	ua
!!! В Харькове вводится комендантский час с 15:00 до 06:00 завтрашнего дня.	ru	ru	ru
ДТП Киев авария парковая дорога большая пробка. Видео Настя спасибо!	ru	ru	ru

Table 5: Randomly selected data points from the selected subset for manual language verification.