EmoBench-UA: A Benchmark Dataset for Emotion Detection in Ukrainian

Anonymous ACL submission

Abstract

While Ukrainian NLP has seen progress in many texts processing tasks, emotion classification remains an underexplored area with no publicly available benchmark to date. In this work, we introduce EmoBench-UA, the first annotated dataset for emotion detection in Ukrainian texts. Our annotation schema is adapted from the previous English-centric works on emotion detection (Mohammad et al., 2018; Mohammad, 2022) guidelines. The dataset was created through crowdsourcing using the Toloka.ai platform ensuring high-quality of the annotation process. Then, we evaluate a range of approaches on the collected dataset, starting from linguistic-based baselines, synthetic data translated from English, to large language models (LLMs). Our findings highlight the challenges of emotion classification in non-mainstream languages like Ukrainian and emphasize the need for further development of Ukrainian-specific models and training resources.

Figure 1: EmoBench-UA is a benchmark of basic emotions—Joy, Anger, Fear, Disgust, Surprise, Sadness, or None—detection in Ukrainian texts.

1 Introduction

Recent trends in natural language processing indicate a shift from predominantly monolingual English-centric approaches toward more inclusive multilingual solutions that support less-resourced and non-mainstream languages. Although cross-lingual transfer techniques—such as Adapter modules (Pfeiffer et al., 2020) or translation from resource-rich languages (Kumar et al., 2023)—have shown promise, the development of high-quality, language-specific datasets remains essential for achieving robust and culturally accurate performance in these settings.

For the Ukrainian language, significant progress has been made in the development of resources for various stylistic classification tasks, such as sentiment analysis (Zalutska et al., 2023) and toxicity detection (Dementieva et al., 2024). However, to the best of our knowledge, no publicly available dataset has yet addressed the task of emotion classification. In this work, we aim to fill this gap through the following contributions:

- We design a robust **crowdsourcing annotation pipeline** for emotion annotation in Ukrainian texts, leveraging the Toloka.ai platform and incorporating quality control mechanisms to ensure high-quality annotations;
- Using this pipeline, we collect **EmoBench-UA**, the first manually annotated benchmark dataset for emotion detection in Ukrainian;
- We evaluate a range of **classification approaches** on the dataset—including linguistic-based baselines, Transformer-based encoders, translation into English, and prompting large language models (LLMs)—to assess task dif-

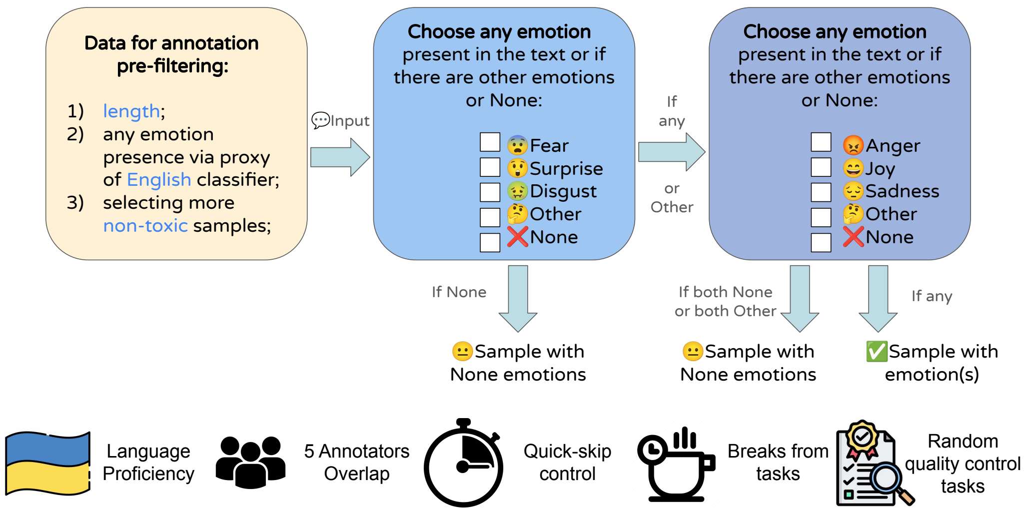


Figure 2: EmoBench-UA Annotation Pipeline: we split the annotation into two tasks to improve annotator focus, and several quality control measures were applied to ensure the high quality of the collected data.

ficulty and provide a comprehensive performance analysis.

We release all the instructions, data, and baselines code fully online for public usage.[1]

2 Related Work

Emotion Detection Datasets and Models As for many NLP tasks, various datasets, lexicon, and approaches in the first order were created for English emotions classification (Mohammad et al., 2018). Then, it was also extended to other popular languages like Spanish, German, and Arabic (Plaza del Arco et al., 2020; Chatterjee et al., 2019; Kumar et al., 2022) and then for some not so mainstream languages like Finish (Öhman et al., 2020). Given the challenges associated with collecting fully annotated emotion datasets across languages, a multilingual emotional lexicon (Mohammad, 2023) which covers 100 languages was proposed by translating the original English resources, offering a practical first step toward facilitating emotion detection in lower-resource scenarios.

At the same time, the importance of developing robust NLP systems for emotion analysis and detection is well recognized (Kusal et al., 2023), especially in socially impactful domains such as customer service, healthcare, and support for minority communities. However, extending emotion detection capabilities uniformly across multiple languages remains a persistent challenge (De Bruyne, 2023). For English and several other languages, a variety of classification methods have been explored, ranging from BiLSTM and BERT-based models (Al-Omari et al., 2020; De Bruyne et al., 2022) to more advanced architectures such as XLM-RoBERTa (Conneau et al., 2020), E5 (Wang et al., 2024a), and multilingual LLMs like BLOOMz (Wang et al., 2024b).

[1]The link will be provided upon the paper acceptance.

Ukrainian Texts Classification Although the availability of training data for classification tasks in Ukrainian remains limited, the research community has made notable strides in mayny NLP tasks. For example, UberText 2.0 (Chaplynskyi, 2023) provides resources for NER tasks, legal document classification, and a wide range of textual sources including news, Wikipedia, and fiction. In addition, the OPUS corpus (Tiedemann, 2012) offers parallel Ukrainian data for cross-lingual applications. Recently, the Spivavtor dataset (Saini et al., 2024) has also been introduced to facilitate instruction-tuning of Ukrainian-focused large language models.

For related classification tasks, resources for sentiment analysis (Zalutska et al., 2023) and toxicity detection (Dementieva et al., 2024) have already been developed for Ukrainian. Additionally, in the domain of abusive language, a bullying detection system for Ukrainian was introduced but based on translated English data (Oliinyk and Matviichuk, 2023). Dementieva et al. (2025) explored various cross-lingual knowledge transfer methods for Ukrainian texts classification, yet emphasized the continued importance of authentic, manually annotated Ukrainian data.

3 EMOBENCH-UA Collection

Here, we present the design of the crowdsourcing collection pipeline, detailing the task setup, annotation guidelines, interface design, and the applied quality control procedures used to obtain EMOBENCH-UA. The overall schema of the pipeline is presented in Figure 2.

3.1 Emotions Classification Objective

In this work, we define emotion recognition as the task of identifying perceived emotions—that is, the emotion that the majority of people would attribute to the speaker based on a given sentence or short text snippet (Mohammad, 2022).

We adopt the set of basic emotions proposed by Ekman et al. (1999), which includes `Joy`, `Fear`, `Anger`, `Sadness`, `Disgust`, and `Surprise`. A single text instance may convey multiple emotions simultaneously creating the **multi-label** classification task. If a text *does not express* any of the listed emotions, then we assign it the label `None`.

3.2 Data Selection for Annotation

As the source data, we selected the publicly available Ukrainian tweets corpus (Bobrovnyk, 2019). Given that social media posts are often rich in emotionally charged content, this corpus serves as a suitable foundation for our annotation task. Since the original collection consists of several hundred thousand tweets, we applied a multi-stage filtering process to both increase the likelihood of emotional content and ensure the feasibility of accurate annotation:

Length First, we applied a length-based filter, discarding texts that were too short (N words < 5), as such samples often consist of hashtags or other non-informative tokens. Similarly, overly long texts (N words ≥ 50) were excluded, as longer sequences tend to obscure the central meaning and make it challenging to accurately identify the expressed emotions.

Toxicity While toxic texts can carry quite strong emotions, to ensure annotators well-being and general appropriateness of our corpus, we filtered out too toxic instances using opensourced toxicity classifier (Dementieva et al., 2024).[2]

Emotional Texts Pre-selection To avoid an excessive imbalance toward emotionless texts, we performed a pre-selection step aimed at identifying texts likely to express emotions. Specifically, we applied the English emotion classifier `DistillRoBERTa-Emo-EN`[3] on translated Ukrainian texts. For this, $10k$ Ukrainian samples, previously filtered by the previous steps, were translated into English using the NLLB model (Costa-jussà et al., 2022)[4]. The emotion predictions from this classifier were then used to select a final set of $5k$ potentially emotionally-relevant texts, which were used for the further annotation.

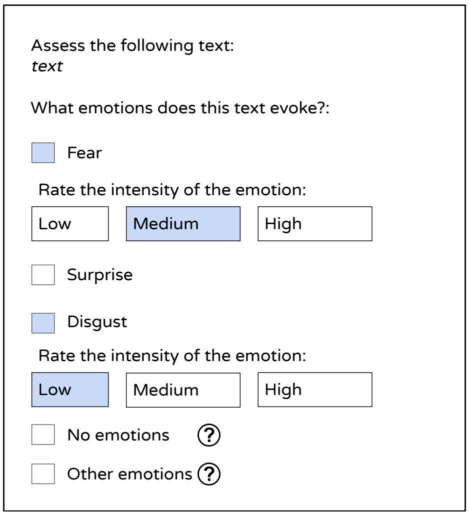


Figure 3: Annotation Interface illustration translated into English.

3.3 Annotation Setup

As emotion classification is quite subjective, we opted to rely on crowdsourcing rather than limiting the annotation process to a small group of expert annotators. For this, we utilized Toloka.ai[5] crowdsourcing platform.

3.3.1 Projects Design

As shown in Figure 2, to reduce cognitive load, we split the annotation process into two separate projects: one focused on *fear*, *surprise*, and *disgust*; the other on *anger*, *joy*, and *sadness*. Annotators could select multiple emotions per sample, with additional options *No emotion* and *Other emotion* provided. If a sample received conflicting annotations across the two projects (e.g., *No emotion* in one and *Other emotion* in the other), it was excluded from the dataset.

[2] https://huggingface.co/ukr-detect/ukr-toxicity-classifier

[3] https://huggingface.co/michellejieli/emotion_text_classifier

[4] https://huggingface.co/facebook/nllb-200-distilled-600M

[5] https://toloka.ai

3.3.2 Instructions & Interface

Before being granted access to the annotation task, potential annotators were provided with detailed instructions, including a description of our aimed emotion detection task and illustrative examples for each emotion. We present the English translation of the introductory part of our instruction text:

> **Instructions**
> Select one or more emotions and their intensity in the text. If there are no emotions in the text or if there are emotions not represented in the list, select the No emotionsother emotions option.

with the full listed Ukrainian version for both projects in Appendix C. The English translation of the interface is presented in Figure 3 with the original Ukrainian interface in Figure 5.

Annotators were instructed to answer a multiple-choice question, allowing them to select one or more emotions for each text instance. Additionally, they were asked to indicate the perceived intensity of the selected emotions. These annotations were also collected and will be included in the final release of EMOBENCH-UA. However, for the purposes of this study in the experiments, we focus exclusively on the binary emotion presence labels.

3.3.3 Annotators Selection

Language Proficiency Toloka platform provided pre-filtering mechanisms to select annotators who had passed official language proficiency tests, serving as an initial screening step. In our scenario, we selected annotators that were proficient in Ukrainian.

Training and Exam Phases Annotators interested in participating first completed an unpaid training phase, where they reviewed detailed instructions and examples with explanations for correct labelling decisions. Following this, annotators were required to pass then an exam, identical in format to the actual labelling tasks, to demonstrate their understanding of the guidelines. Successful candidates gained access to the main assignments.

3.3.4 Quality Control

To ensure high-quality annotations, we implemented several automated checks. Annotators were permanently banned if they submitted the last three task pages in under 15 seconds each, indicating low engagement. A one-day ban was triggered if three consecutive pages were skipped. To prevent fatigue, annotators were asked to take a 30-minute break after completing 25 consecutive pages. Additionally, control tasks were randomly injected; if the accuracy on these within the last 10 pages fell below 40%, the annotator was temporarily banned and required to retake the training.

To ensure the reliability of the annotations, each text instance was labeled independently by 5 annotators. The final emotion labels were determined through majority voting with an estimated confidence score. Only instances with a confidence score $\geq 90\%$ were included to the final dataset.

3.3.5 Annotators Well-Being

We aimed to design a fair, transparent, and user-friendly crowdsourcing project.

Fair Compensation Payment rates were set to balance grant funding constraints with fair wages, aligning with Ukraine's minimum hourly wage at the time of labelling (**1.12 USD/hour**). Annotators received **0.05 USD** per page with possibility to complete at least 20 assignment per hour. The overall spending of the whole project resulted in **500 USD**.

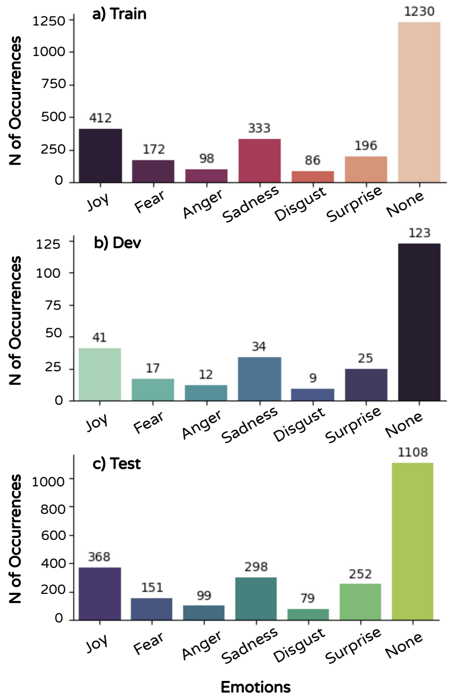
Positive Project Ratings Toloka provided annotators with tools[6] to rate project fairness in terms of payment, task design, and organizer responsiveness. Our projects received high ratings: **4.80/5.00** for the Training Project and **4.90/5.00** for the Main Project.

4 EMOBENCH-UA

After filtering out low-confidence and ambiguous samples from the annotation results, we obtained a final EMOBENCH-UA of 4949 labelled instances (145 samples were dropped due to label conflicts). Krippendorff's alpha agreement score was 0.85. Then, we partitioned the dataset into fixed train/development/test subsets following a 50/5/45% split ratio. An overview of the label distribution across these subsets is presented in Figure 4a. The dataset examples can be found in Appendix D.

We were able to collect at least one hundred, and in some cases several hundred, instances for each emotion category. Nevertheless, the dataset remains imbalanced, with `Joy` and `Sadness` being the most prevalent emotions among the labeled samples, alongside a substantial portion of texts assigned the `None` label. Such imbalance is a common characteristic of emotion detection datasets,

[6] https://toloka.ai/docs/guide/project_rating_stat



(a) Distribution of Labels

(b) Keywords per Emotion

Figure 4: EmoBench-UA statistics per sets and emotions.

reflecting the natural distribution of emotions in real-world text and contributing to the overall complexity of the task. Additionally, in Figure 4b, we provide a closer analysis of the collected emotional data by extracting the top-10 keywords for each emotion label (lemmatization done using the spacy[7] library). The resulting keywords reveal clear and intuitive associations with the corresponding emotional categories, further confirming the quality and relevance of the annotated texts.

5 Models

We test various types on models on our collected dataset: (i) linguistic-based approaches; (ii) Transformer-based encoders; (iii) LLMs prompting for classification. Then, we also did an ablation study with synthetic training Ukrainian data acquisition via translation from English. The details of models hyperparameters can be found in Appendix F.

5.1 Linguistic-based Approaches

Even with current advances in NLP, linguistic-based approaches based on statistics of the training set can be quite a strong and resource-efficient baseline for stylistic texts classification like sentiment (Brauwers and Frasincar, 2023) or formality (Dementieva et al., 2023).

Keywords Based We used the train part of our dataset to extract *natural* keywords per emotion as shown in Figure 4b. We used `spacy` for lemmatization extracting top-20 words per emotion.

Logistic Regression Firstly, we embed our texts with `CountVectorizer` into `td-idf` features. Then, we fine-tuned Logistic Regressions classifier on the training part of our dataset.

Random Forest The same as for logistic regression, we fine-tune Random Forest classifier with 100 decision trees on `td-idf` training features.

5.2 Transformer-based Encoder

Then, we take the next generation of classification models based on the Transformers (Vaswani et al., 2017) encoders. For each model type, we evaluate multiple versions varying in model size.

BERT Firstly, we used `mBERT`[8] (Devlin et al., 2018) as it contains Ukrainian in the pre-trained data. We additionally experimented with a compact variant—`Geotrend-BERT`[9]—of mBERT where the

[7] https://spacy.io/models/uk

[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

[9] https://huggingface.co/Geotrend/bert-base-uk-cased

vocabulary and embeddings were specifically refined to retain only Ukrainian (Abdaoui et al., 2020).

RoBERTa As an extension of BERT-alike models, we used several versions of RoBERTa-alike models (Conneau et al., 2019) as it shown previously promising results in Ukrainian texts classification (Dementieva et al., 2025):

- `XLM-RoBERTa`: `base`[10] and `large`[11] instances;
- Ukrainian-specific pre-trained monolingual RoBERTa: `UKR-RoBERTa-base`[12];
- additionally fine-tuned on sentiment classification task on Twitter data `Twitter-XLM-RoBERTa base`[13] (Barbieri et al., 2022);
- finally, we tested `Glot500-base`[14] model (Imani et al., 2023) that extended multilingual RoBERTa to 500 languages.

LaBSe Another multilingual embedding model covering 109 languages including Ukrainian: `LaBSe`[15] (Feng et al., 2022).

E5 Finally, we utilized the more recent `multilingual-e5` embeddings (Wang et al., 2024a): `base`[16] and `large`[17] variants.

5.3 LLMs prompting

To test models based on another methodology, we also tried out various modern LLMs on our benchmark dataset transforming our classification task into the `text-to-text` generation one. While Ukrainian is not always explicitly present in the pre-training data reports, the emerging abilities of LLMs already showed promising results in handling new languages (Wei et al., 2022) including Ukrainian (Dementieva et al., 2025). However, we also utilize more recent LLMs dedicated to European languages, including Ukrainian. We used two types of prompts—instructions in English and in Ukrainian—that are fully listed in Appendix E.

We tested several families of LLMs with variants in terms of version and sizes. We chose mostly instruction tuned instances as they supposedly perform more precise for classification tasks:

EuroLLM The recent initative introduced in (Martins et al., 2024) has an aim to develop high-quality LLMs for European languages with Ukrainian definitely included. We selected `EuroLLM-1.7B`[18] variant for our experiments.

Mistral We used several version of Mistral-family models (Jiang et al., 2023)—`Mistral-7B`[19] and `Mixtral-8x7B`.[20] The models cards do not mention explicitly Ukrainian and other languages, however Mistral showed promising results in Ukrainian texts classification tasks (Dementieva et al., 2025).

LLaMa3 The LLaMa model (AI@Meta, 2024) card as well does not stated Ukrainian explicitly, however, encourages research in usage of the model in various multilingual tasks. Thus, we tested the `Llama-3-8B`[21] and `Llama-3.3-70B`[22] variants.

DeepSeek Finally, we tested one of the recent top performing models in reasoning—DeepSeek (DeepSeek-AI et al., 2025) with `DeepSeek-R1-Qwen`[23], `deepseek-ai/DeepSeek-R1-Llama`[24], and `DeepSeek-V3`[25] variants. The situation of the Ukrainian language presence in the models is the same as for Mistral and LLaMa—DeepSeek was heavily optimized for English and Chinese, however, the authors encourage to try it for other languages.

5.3.1 Translation & Synthetic Data

Additionally, we also experimented with transnational setups to imitate various low-resource scenarios: (i) translation in `ukr`→`en` direction; (ii) translation in `en`→`ukr` direction.

Emotion Lexicon In addition to natural Ukrainian lexicon extracted from our data, we also experimented with the already collected and translated from English *synthetic* Ukrainian emotions lexicon from (Mohammad, 2023).

Backtranslation Then, we imitated the scenario if we have already fine-tuned English emotion detection model—i.e. `DistillRoBERTa-Emo-EN`[26]—

[10] https://huggingface.co/FacebookAI/xlm-roberta-base
[11] https://huggingface.co/FacebookAI/xlm-roberta-large
[12] https://huggingface.co/youscan/ukr-roberta-base
[13] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[14] https://huggingface.co/cis-lmu/glot500-base
[15] https://huggingface.co/sentence-transformers/LaBSE
[16] https://huggingface.co/intfloat/multilingual-e5-base
[17] https://huggingface.co/intfloat/multilingual-e5-large
[18] https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct
[19] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[20] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[21] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[22] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[23] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[24] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
[25] https://huggingface.co/deepseek-ai/DeepSeek-V3
[26] https://huggingface.co/michellejieli/emotion_text_classifier

	Joy	Fear	Anger	Sadness	Disgust	Surprise	None	Pr	Re	F1
Linguistic-based Approaches										
Keywords	0.30	0.15	0.08	0.21	0.10	0.15	0.25	0.24	0.24	0.22
Logistic Regression	**0.64**	**0.72**	**0.49**	**0.59**	**0.49**	**0.61**	0.67	0.51	**0.22**	**0.29**
Random Forest	0.61	0.69	**0.49**	**0.59**	**0.49**	0.60	**0.68**	**0.58**	0.21	0.27
Backtranslation										
DistillRoBERTa-Emo-EN	0.56	0.55	0.31	0.52	0.23	0.47	0.55	0.40	0.61	0.45
Transformer-based Encoders										
LaBSe	0.67	0.73	0.30	0.65	0.33	0.54	0.80	0.57	0.59	0.57
Geotrend-BERT	**0.58**	**0.59**	**0.08**	**0.50**	**0.11**	**0.40**	**0.73**	**0.46**	**0.43**	**0.43**
mBERT	0.46	0.24	0.01	0.45	0.02	0.33	0.73	0.33	0.33	0.32
UKR-RoBERTa Base	0.65	0.58	0.14	0.50	**0.21**	0.49	0.74	0.51	0.45	0.47
XLM-RoBERTa Base	0.61	0.31	0.00	0.33	0.01	0.19	0.75	0.33	0.31	0.31
XLM-RoBERTa Large	**0.73**	**0.79**	**0.20**	**0.68**	0.00	**0.60**	**0.80**	0.52	**0.58**	**0.54**
Twitter-XLM-RoBERTa	0.72	0.76	0.13	0.64	0.07	0.54	0.79	**0.66**	0.51	0.52
Glot500 Base	0.01	0.02	0.03	0.18	0.00	0.01	0.64	0.24	0.19	0.13
Multilingual-E5 Base	0.71	0.73	0.01	0.52	0.00	0.50	0.77	0.49	0.45	0.46
Multilingual-E5 Large	**0.73**	**0.81**	**0.31**	**0.69**	**0.35**	**0.60**	**0.81**	**0.65**	**0.62**	**0.62**
LLMs Prompting										
EuroLLM-1.7B (ENG)	**0.46**	**0.31**	**0.15**	**0.37**	**0.18**	0.09	**0.28**	**0.26**	**0.38**	**0.26**
EuroLLM-1.7B (UKR)	0.38	0.30	0.11	0.27	0.10	**0.11**	0.25	0.25	0.24	0.22
Mistral-7B (ENG)	0.52	**0.58**	**0.33**	**0.49**	**0.32**	**0.37**	**0.52**	**0.37**	**0.73**	**0.45**
Mistral-7B (UKR)	**0.55**	0.37	0.28	0.47	0.19	0.24	0.33	0.32	0.71	0.35
Mixtral-8x7B (ENG)	**0.49**	**0.37**	**0.34**	**0.51**	**0.25**	**0.25**	0.66	**0.32**	**0.74**	**0.41**
Mixtral-8x7B (UKR)	0.48	0.35	0.19	0.47	0.21	0.22	**0.71**	0.27	0.73	0.37
LLaMA 3 8B (ENG)	**0.56**	0.65	**0.36**	**0.54**	**0.29**	**0.25**	**0.39**	**0.43**	**0.56**	**0.43**
LLaMA 3 8B (UKR)	0.30	**0.67**	0.29	0.45	0.15	**0.25**	0.10	0.38	0.53	0.31
LLaMA 3.3 70B (ENG)	**0.64**	0.63	**0.47**	0.62	**0.26**	0.32	**0.43**	0.44	**0.79**	**0.48**
LLaMA 3.3 70B (UKR)	0.58	**0.68**	0.34	**0.71**	0.18	**0.33**	0.36	**0.45**	0.64	0.46
DeepSeek-R1-Qwen (ENG)	0.63	0.61	**0.43**	**0.64**	**0.45**	**0.46**	0.60	**0.48**	**0.75**	**0.55**
DeepSeek-R1-Qwen (UKR)	**0.68**	**0.66**	0.40	0.57	0.29	0.38	**0.68**	0.46	0.66	0.52
DeepSeek-R1-LLaMA (ENG)	**0.67**	**0.69**	**0.49**	**0.71**	**0.52**	0.47	0.67	**0.54**	**0.72**	**0.60**
DeepSeek-R1-LLaMA (UKR)	**0.67**	0.64	0.45	0.69	0.33	**0.51**	**0.69**	0.51	0.69	0.57
DeepSeek-V3 (ENG)	**0.73**	0.74	0.60	**0.72**	**0.57**	0.41	**0.78**	**0.60**	0.72	**0.65**
DeepSeek-V3 (UKR)	0.71	0.66	**0.61**	**0.72**	0.48	**0.42**	0.71	0.54	**0.81**	0.62

Table 1: EmoBench-UA test set results of various models types per emotion and overall. In **bold**, we denote the best results per column within model type. In orange we highlight the top results per column.

so then we can translate Ukrainian inputs into English to obtain the labels.

Synthetic Training Data via Translation To not rely on the translation everytime at inference, we can also translate the whole English training corpus (Muhammad et al., 2025) into Ukrainian and then used it as Ukrainian training data.

For translation in all scenarios, we utilized NLLB[27] model (Costa-jussà et al., 2022).

6 Results

The results of models evaluation on the test part of our novel EmoBench-UA dataset on the **binary multi-label classification** task are presented in the Table 1. We report **F1 score** per each emotion; for overall results, we report Precision, Recall, and **macro-averaged F1-score**. Also, we provide the confusion matrices for the top performing models in Appendix G.

Linguistic-based Approaches While the linguistic-based models rely on relatively simple statistical representations of the text, they demonstrate competitive performance. The keyword-based approach, however, yielded lower results, which is expected given that emotion detection often relies on understanding contextual collocations and multi-word expressions rather than isolated words. In contrast, both logistic regression and random forest models performed on par with several base encoder models and, in some cases, even outperformed certain LLMs. Although these models did not achieve the highest overall F1-macro scores, they showed strong precision but struggled with recall.

Backtranslation The approach of leveraging an English-based classifier as a proxy demonstrated

[27] https://huggingface.co/facebook/nllb-200-distilled-600M

	Joy	Fear	Anger	Sadness	Surprise	None	Pr	Re	F1
Keywords UK	**0.30**	**0.15**	**0.08**	**0.21**	**0.15**	**0.25**	**0.27**	**0.25**	**0.26**
Keywords EN	0.17	0.05	0.01	0.18	0.08	0.11	0.15	0.01	0.10
UKR-RoBERTa-base UK	**0.65**	**0.58**	0.14	**0.50**	**0.49**	**0.74**	**0.56**	**0.49**	**0.52**
UKR-RoBERTa-base EN	0.53	0.24	**0.19**	0.30	0.31	0.60	0.32	0.42	0.36
mBERT UK	**0.46**	**0.24**	0.00	**0.45**	0.33	**0.73**	**0.38**	**0.38**	**0.37**
mBERT EN	0.38	0.12	**0.12**	0.31	**0.31**	0.55	0.31	0.30	0.30
LaBSe UK	**0.67**	**0.73**	**0.30**	**0.65**	**0.54**	**0.80**	**0.59**	**0.65**	**0.62**
LaBSE EN	0.60	0.41	0.22	0.39	0.30	0.64	0.44	0.43	0.43
XLM-RoBERTa Large UK	**0.73**	**0.79**	**0.20**	**0.68**	**0.60**	**0.80**	**0.61**	**0.68**	**0.63**
XLM-RoBERTa Large EN	0.50	0.34	0.15	0.47	0.24	0.53	0.33	0.45	0.37
Twitter-XLM-RoBERTa UK	**0.72**	**0.76**	0.13	**0.64**	**0.54**	**0.79**	**0.60**	**0.59**	**0.60**
Twitter-XLM-RoBERTa EN	0.62	0.26	**0.21**	0.52	0.44	0.62	0.42	0.47	0.44
Multilingual-E5 Large UK	**0.73**	**0.81**	**0.31**	**0.69**	**0.60**	**0.81**	**0.65**	**0.68**	**0.66**
Multilingual-E5 Large EN	0.61	0.26	0.22	0.36	0.23	0.56	0.36	0.41	0.37

Table 2: EmoBench-UA test set results of comparison natural `UK` vs synthetic `EN` training data. In **bold**, we denote the best results per column within model type. As the English dataset does not contain `Disgust` label, we fine-tuned all models types without it for this experiment.

competitive performance as well. Notably, it achieved one of the highest scores for the `Anger` category, where many other models struggled. Although its precision was lower compared to even the linguistic-based methods, it consistently delivered substantially higher recall. Thus, it can be quite a good basline for Ukrainian emotional texts detection.

Transformer-based Encoders Among the range of tested BERT- and RoBERTa-based models, the Ukrainian-specific encoders, `Geotrend-BERT` and `UKR-RoBERTa Base`, significantly outperformed `mBERT`, `Glot500`, and `XLM-RoBERTa-base`, highlighting the importance of monolingual, Ukrainian-specific encoders. At the same time, the multilingual `LaBSE` model outperformed Ukrainian-specific models. Within the RoBERTa-like family, `XLM-RoBERTa-large` and `Twitter-XLM-RoBERTa` achieved the strongest results, although both struggled with the `Anger` and `Disgust`. Finally, the best-performing encoder was **`Multilingual-E5-Large`**, with a good balance of Precision and Recall.

LLMs Across all model families, we observe a consistent trend of slightly improved performance when models are prompted in English rather than Ukrainian. Surprisingly, `EuroLLM` underperformed, yielding results even lower than the linguistics-based baselines. Other LLMs delivered scores comparable to encoder-based models, outperforming them in the `Anger` and `Disgust` classes. While all LLMs demonstrated lower Precision compared to the best encoders, they consistently achieved higher Recall. Notably, **`DeepSeek-V3`** handled the emotion detection task in Ukrainian with the highest scores. However, the overall performance gains over `Multilingual-E5-Large` remain minimal, raising a question regarding the efficiency and responsible usage of such large models.

Natural vs Translated Data From Table 2, we observe that models trained on the original Ukrainian data consistently outperform their English-tuned counterparts. However, the latter in some cases achieve higher scores for the `Anger` class, suggesting—in line with previous observations with the models containing knowledge of English—that English data could be a valuable for augmenting Ukrainian samples for it.

7 Conclusion

We introduced EmoBench-UA—the first manually annotated dataset for emotion detection in Ukrainian texts. The proposed pipeline combines data preprocessing with a two-stage annotation procedure, incorporating multiple quality control measurements to ensure the high quality annotation. We benchmarked a wide range of approaches for the multi-label emotion classification task, demonstrating that although the latest LLMs, such as DeepSeek, achieved the strongest results, more efficient encoder-based models perform competitively. We hope this work encourages further research on Ukrainian-specific emotion detectors, including ensemble strategies and augmentation with English-based resources.

Limitations

While this work introduces EMOBENCH-UA as a valuable benchmark for emotion detection in Ukrainian texts, we acknowledge several limitations worth addressing and exploring in future research.

Emotions Labels The current dataset is restricted to the recognition of basic emotions. More nuanced or implicit emotional states, which often arise in real-world communication, remain outside the scope of this release.

Another challenge is the interpretation of the `None` label, which can reflect both an absence of emotion or still can be a holder for other emotions rather then listed basic ones. Distinguishing between these two cases is non-trivial and requires deeper investigation.

Emojis as Keywords The role of non-verbal cues—in particular, the presence of emojis in social media texts—has not been systematically investigated in this work. Emojis can often serve as strong emotion indicators, and future experiments could benefit from incorporating emoji-aware detectors.

Crowdsourcing Platform Additionally, while the annotation process was performed on a specific crowdsourcing platform—Toloka.ai—we believe that the design of the annotation pipeline is platform-agnostic as annotation guidelines and quality control measures are openly available.

Annotators Overlap Although each instance in the dataset was annotated by five independent annotators, emotions are still highly subjective and culturally sensitive. Increasing annotator overlap, as well as ensuring broader demographic diversity—i.e. Ukrainian speakers from various regions of the country—could further improve label robustness.

Detectors Design This study focused on evaluating one representative model per classifier type. Future work could explore ensemble methods or hybrid architectures, which have the potential to further enhance performance.

Hyperparameters Lastly, hyperparameter optimization was explored in a limited setup. More systematic tuning, particularly for prompting strategies (e.g., temperature settings) and fine-tuning, is likely to yield additional improvements.

8 Ethics Statement

We also consider several ethical implications of our work.

During data collection, we made our best to ensure that all contributors were fairly compensated. Clear guidelines and examples were provided to reduce potential ambiguity or emotional strain on the annotators.

All texts in the dataset originate from publicly available sources and were anonymized with totally removed links and any users mentioning to avoid the disclosure of personal or sensitive information. Nonetheless, since the source data comes from social media, there remains a potential for indirect identification through unique expressions or context. We encourage future users of the dataset to handle the material responsibly.

Given the subjective nature of emotions and their cultural grounding, we acknowledge that both annotation and model predictions may reflect current social and cultural biases. This is a general limitation for emotion or other style recognition datasets. We advise the stakeholders of the potential applications to additionally cross-check the models and data for their specific use-cases with corresponding to context adjustments.

Finally, we openly release the annotation guidelines for transparency and reproducibility and encourage future work to continue contribute with various data, including emotions detection, for underrepresented languages.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed: 2024-12-14.

Hani Al-Omari, Malak A Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232. IEEE.

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.

Kateryna Bobrovnyk. 2019. Automated building and analysis of ukrainian twitter corpus for toxic text detection. In *COLINS 2019. Volume II: Workshop*.

Gianni Brauwers and Flavius Frasincar. 2023. A survey on aspect-based sentiment classification. *ACM Comput. Surv.*, 55(4):65:1–65:37.

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 458–466, Toronto, Canada. Association for Computational Linguistics.

Luna De Bruyne, Pranaydeep Singh, Orphee De Clercq, Els Lefever, and Veronique Hoste. 2022. How language-dependent is emotion detection? evidence from multilingual BERT. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 76–85, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2023. Detecting text formality: A study of text classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024. Toxicity classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Valeriia Khylenko, and Georg Groh. 2025. Cross-lingual text classification transfer: The case of Ukrainian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1451–1464, Abu Dhabi, UAE. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Paul Ekman, Tim Dalgleish, and M Power. 1999. Basic emotions. *San Francisco, USA*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Puneet Kumar, Kshitij Pathania, and Balasubramanian Raman. 2023. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data. *Appl. Intell.*, 53(9):10096–10113.

Shivani Kumar, Anubhav Shrimal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowl. Based Syst.*, 240:108112.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Rahul Vora, and Ilias O. Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artif. Intell. Rev.*, 56(12):15129–15215.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *CoRR*, abs/2409.16235.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

V Oliinyk and I Matviichuk. 2023. Low-resource text classification using cross-lingual models for bullying detection in the ukrainian language. *Adaptive systems of automatic control: interdepartmental scientific and technical collection, 2023, 1 (42).*

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Aman Saini, Artem Chernodub, Vipul Raheja, and Vivek Kulkarni. 2024. Spivavtor: An instruction tuned Ukrainian text editing model. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 95–108, Torino, Italia. ELRA and ICCL.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 text embeddings: A technical report. *CoRR*, abs/2402.05672.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024b. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Olha Zalutska, Maryna Molchanova, Olena Sobko, Olexander Mazurets, Oleksandr Pasichnyk, Olexander Barmak, and Iurii Krak. 2023. Method for sentiment analysis of ukrainian-language reviews in e-commerce using roberta neural network. In *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems. Volume I: Machine Learning Workshop, Kharkiv, Ukraine, April 20-21, 2023*, volume 3387 of *CEUR Workshop Proceedings*, pages 344–356. CEUR-WS.org.

A Licensing of Resources

Below is an overview of the licenses associated with each resource used in this work (Table 3).

Resource	License	Homepage
Our dataset	CC BY 4.0	*will be provided upon acceptance*
Ukrainian Tweets Dataset	CC BY 4.0	https://ena.lpnu.ua:8443/server/api/core/bitstreams/c4c645c1-f465-4895-98dd-765f862cf186/content
Ukrainian Toxicity Classifier	OpenRail++	https://huggingface.co/ukr-detect
Emotion Lexicon	The lexicon is made freely available for research, and has been commercially licensed to companies for a small fee	https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
mBERT	Apache-2.0	https://huggingface.co/google-bert
Geotrend-BERT	Apache-2.0	https://huggingface.co/Geotrend/bert-base-uk-cased
XLM-RoBERTa	MIT	https://huggingface.co/FacebookAI
UKR-RoBERTa	MIT	https://github.com/youscan/language-models
Twitter-XLM-RoBERTa	Apache-2.0	https://aclanthology.org/2022.lrec-1.27
Glot500	CC BY 4.0	https://aclanthology.org/2023.acl-long.61
LaBSE	Apache-2.0	https://huggingface.co/sentence-transformers/LaBSE
e5	MIT	https://huggingface.co/intfloat
Mistral7B	Apache-2.0	https://huggingface.co/mistralai
Mixstral8x7B	Apache-2.0	https://huggingface.co/mistralai
EuroLLM	Apache-2.0	https://huggingface.co/utter-project/EuroLLM-1.7B-Instruct
LLaMa3	llama3	https://huggingface.co/meta-llama
DeepSeek	MIT	https://huggingface.co/collections/deepseek-ai/deepseek-r1-678e1e131c0169c0bc89728d
NLLB	CC BY NC 4.0	https://huggingface.co/facebook/nllb-200-distilled-600M

Table 3: Overview of the licenses associated with each resource.

The licenses associated with the models and datasets utilized in this study are consistent with the intended use of conducting academic research on approaching various NLP application for positive impact.

B Usage of AI Assistants

During this study, AI assistant was utilized in the writing process. ChatGPT was employed for paraphrasing and improving clarity throughout the paper's formulation. We also utilized DeepL[28] to translate the examples in Ukrainian into English followed by the human manual check and adjustments.

[28] https://www.deepl.com

C Instructions & Interface

C.1 Ukrainian (original)

In this section, we provide the Instructions for both annotation projects as well as interface in Ukrainian.

Main Instructions for the First Project: Fair, Surprise, Disgust

Виберіть одну або кілька емоцій та їх інтенсивність у тексті. Якщо в тексті немає ніяких емоцій або є емоції не представлені в списку виберіть варіант - ”Немає емоцій / інші емоції”.

Приклади

Страх
Низька проява
    Що, як це ніколи не закінчиться?

Нормальна проява
    Мені дуже страшно залишатися тут одному...

Інтесивна проява
    Боже, який це жах і як же це страшно!!!

Здивування
Низька проява
    Це було несподівано

Нормальна проява
    Це вражає! Я в захваті!

Інтесивна проява
Вауyy, який неймовірний поворот подій!!!

Огида
Низька проява
    Щось мене трохи нудить від цього запаху.

Нормальна проява
    Фу, це просто огидно!

Інтесивна проява
    Мені стає погано від однієї лише думки про це

Приклади з декількома емоціями
    Ти ще куриш на ходу в таку погоду. – здивування, огида
    Я боюсь, що це все виявиться п’яними розмовами. – огида, страх
    Я не можу повірити, що це дійсно сталося! Це так страшно! – здивування, страх
    Як це можливо? Я боюся навіть уявити, що буде далі! – здивування, страх
    Я не можу повірити, що хтось може їсти таке! Це жахливо! – огида, здивування

Немає емоцій / інші емоції

Немає емоцій

Сьогодні вранці йшов дощ.
Він прочитав книгу за два дні.
Я бачив її вчора на вулиці.

Інші емоції

Я вкрай роздратований цим безладом!
Моє серце розривається від болю :(
Нарешті ми це зробили :):) я просто на сьомому небі від щастя!

Main Instructions for the Second Project: Joy, Sadness, Anger

Виберіть одну або кілька емоцій та їх інтенсивність у тексті. Якщо в тексті немає ніяких емоцій або є емоції не представлені в списку виберіть варіант - "Немає емоцій / інші емоції".

Приклад

Радість
Низька проява

Твоя усмішка робить мій день.

Нормальна проява

Це один з найкращих подарунків, який я коли-небудь отримував.
Це було дуже весело та чудово, наш відпочинок вдався!!

Інтесивна проява

Нарешті ми це зробили!!!!! я просто на сьомому небі від щастя!
Ми виграли!!! :):) Я не можу повірити, що це сталося!

Сум
Низька проява

Цей день був важкий для мене.

Нормальна проява

Я не можу повірити, що це сталося з нами...

Інтесивна проява

Моє серце розривається від болю :((

Гнів
Низька проява

Це мене бісить

Нормальна проява

Я вкрай роздратований цим безладом!

Інтесивна проява

Це абсолютно неприпустимо!!!

Приклади з декількома емоціями

Нарешті ми знайшли ідеальне місце для відпочинку, і це навіть краще, ніж я міг собі уявити! – радість, здивування

Вау, як неочікувано, це найкращий подарунок, який я коли-небудь отримував! – радість, здивування

Мені приємно, що ти прийшов, але ти капець як запізнився!!! – радість, гнів

Мені важко прийняти, що все закінчилося саме так, і я злюся на тебе за це. – сум, гнів

Це так прикро і гнітюче, що наші відносини закінчилися через твою брехню! – гнів, сум

Немає емоцій / інші емоції

Немає емоцій

Сьогодні вранці йшов дощ.

Він прочитав книгу за два дні.

Я бачив її вчора на вулиці.

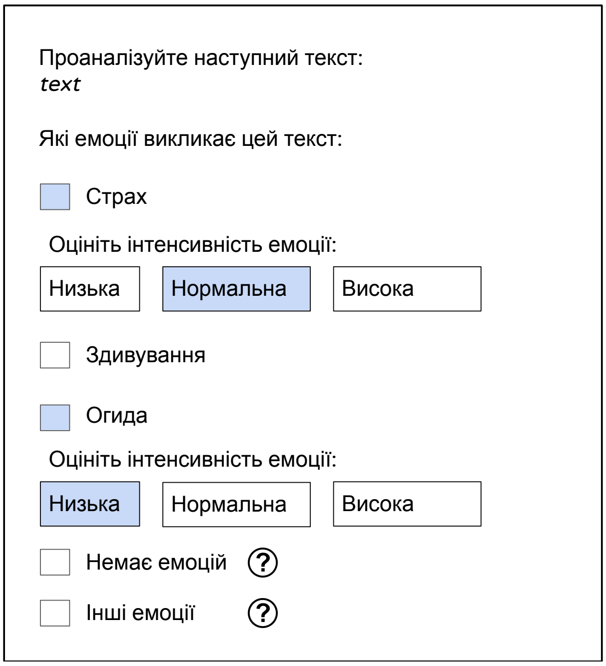


Figure 5: Annotation Interface illustration in original Ukrainian.

C.2 English (translated)

Main Instructions for the First Project: Fair, Surprise, Disgust

Select one or more emotions and their intensity in the text. If there are no emotions in the text or if there are emotions not represented in the list, select the No emotions / other emotions option.

Examples

Fear
Low
What if it never ends?

Normal
I am very scared to stay here alone...

High
My God, what a horror and how scary it is!!!

Surprise
Low
It was unexpected

Normal
It's amazing! I'm thrilled!

High
Wow, what an incredible turn of events!!!

Disgust
Low
This smell makes me a little nauseous.

Normal
Ew, that's just disgusting!

High
I feel sick just thinking about it

Examples with multiple emotions
You're still smoking on the go in this weather. - surprise, disgust
I'm afraid it will all turn out to be drunken talk. - disgust, fear
I can't believe this really happened! It's so scary! - surprise, fear
How is this possible? I'm afraid to even imagine what will happen next! - surprise, fear
I can't believe someone would eat that! It's horrible!" - disgust, surprise

No emotions / other emotions

No emotions
This morning it was raining.
He read the book in two days.
I saw her yesterday on the street.

Other emotions
I am extremely annoyed with this mess!
My heart is breaking with pain :(
We finally did it :):) I'm just over the moon!

Main Instructions for the Second Project: Joy, Sadness, Anger

Select one or more emotions and their intensity in the text. If there are no emotions in the text or if there are emotions not represented in the list, select the No emotions / other emotions option.

Example

Joy

Low

Your smile makes my day.

Normal

This is one of the best gifts I have ever received.
It was very fun and wonderful, our vacation was a success!!!

High

We finally did it!!!!! I'm just over the moon
We won!!! :):) I can't believe it happened!

Sadness

Low

It was a hard day for me.

Normal

I can't believe this happened to us...

High

My heart is breaking with pain :((

Anger

Low

It pisses me off

Normal

I am extremely annoyed with this mess!

High

This is absolutely unacceptable!!!

Examples with multiple emotions

We finally found the perfect place to stay, and it's even better than I could have imagined! - joy, surprise
Wow, how unexpected, this is the best gift I've ever received! - joy, surprise
I'm glad you came, but you're so damn late! - joy, anger
It's hard for me to accept that it ended this way, and I'm angry with you for it. - sadness, anger
It's so sad and depressing that our relationship ended because of your lies! - anger, sadness

No emotions / other emotions

This morning it was raining.

He read the book in two days.

I saw her yesterday on the street.

D EmoBench-UA Samples Examples

Emotion	Data Examples	Intensity
Joy	То так мило і гарно. *It's so nice and beautiful.*	Low
	вже майже час слухаю співи, це справді шикарно*-* *I've been listening to the singing for almost an hour now, it's really great*-**	Medium
	І найголовніше, з Новим роком, пташки!!! *And most importantly, Happy New Year, birds!!!*	High
Fear	Бо я прокинулась, глянула в дзеркало і злякалась. *Because I woke up, looked in the mirror, and got scared.*	Low
	Поспала годинку і почали снитись жахіття :(*I slept for an hour and started having nightmares :(*	Medium
	А в мене руки трусяться) !!! *And my hands are shaking) !!!*	High
Anger	Спілкувалась я з деякими, і от бісить і всьо тут *I talked to some of them, and this is what makes me angry*	Low
	Ставте крапку, мати вашу я знав! *Put a stop to it, I knew your mother, damn it!*	Medium
	Просто чoooрт, ну якого я такий ідіот?!?! *Why am I such an idiot?!?!*	High
Sadness	Але за дітками і їхніми обнімашки скучила. *But I missed my children and their hugs.*	Low
	Не виходить смачний чай:// вкотре *I can't make delicious tea:// once again*	Medium
	Але вона не живе зі мною ((((і я сумую. *But she doesn't live with me ((((and I miss her.*	High
Disgust	В Києві душно, брудно, нудно і нема чим дихати. *Kyiv is stuffy, dirty, boring, and there is no air to breathe.*	Low
	Гірлянди там галімі, а свічки смердючі. *The lights are crappy, and the candles are stinky.*	Medium
	відповідь очевидна – там лайно, фууу!! *the answer is obvious - it's shit, ewww!*	High
Surprise	не може бути, а чому? *it can't be, and why?*	Low
	а шо це, шоце? я шось не бачила такого? *what's this, what's this? I haven't seen anything like it?*	Medium
	а я то думала...он воно що!! *and here I was thinking... but that's it!!!*	High
None	Знову вертоліт над #lviv *Helicopter over #lviv again*	
	поки що не хочу дітей *i don't want children yet*	
	Гуляю собі галицьким селом тихою дорогою. *I'm walking along a quiet road in one Halychyna village.*	

Table 4: EmoBench-UA dataset examples per each emotions.

E LLMs Prompts for Emotions Classification

Here, we provide exact prompts used for LLMs prompting for emotion classification task in Ukrianian texts. We used two types of prompts: instructions in English and instructions in Ukrianian.

Prompt with Instructions in English

Evaluate whether the following text conveys any of the following emotions: joy, fear, anger, sadness, disgust, surprise.
If the text does not have any emotion, answer neutral.
One text can have multiple emotions.
Think step by step before you answer. Answer only with the name of the emotions, separated by comma.

Examples:

Text: Але, божечко, як добре вдома.
Answer: joy

Text: Я в п'ятницю признавалась в коханні і мене відшили!
Answer: sadness

Text: Починаю серйозно хвилюватись за котика.
Answer: fear

Text: Я тебе ненавиджу, п'яна як може бути!
Answer: anger

Text: Тут смердить і мальчіки з синім волоссям п'ють.
Answer: disgust

Text: А що, цей канал досі існує?
Answer: surprise

Text: Хочу вже наводити порядок в новому домі.
Answer: neutral

Text: input
Answer:

Prompt with Instructions in Ukrainian

Оціни, чи передає текст будь-які з цих емоцій: радість, злість, страх, сум, здивування, огида.
Якщо в тексті немає емоцій, відповідай нейтральна.
Один текст може викликати багато емоцій.
Думай крок за кроком, перш ніж відповідати. Відповідай тільки назвами емоцій розділених комою.

Приклади:

Тект: Але, божечко, як добре вдома.
Відповідь: радість

Тект: Я в п'ятницю признавалась в коханні і мене відшили!
Відповідь: сум

Тект: Починаю серйозно хвилюватись за котика.
Відповідь: страх

Тект: Я тебе ненавиджу, п'яна як може бути!
Відповідь: злість

Тект: Тут смердить і мальчіки з синім волоссям п'ють.
Відповідь: огида

Тект: А що, цей канал досі існує?
Відповідь: здивування

Тект: Хочу вже наводити порядок в новому домі.
Відповідь: нейтральна

Текст: input
Відповідь:

F Model hyperparameters

Here, we report the hyperparameters details for the utilized models.

Table 5 reports the tuned `learning rates` per each Transformer-encoder based models. Within all models, we used `batch size` 64, 50 `epochs` with `early stopping callback` 3 according to the accuracy of the evaluation.

For LLMs, for generation, we used default hyperparameters per model with no additional changes.

Model	**Learn. rate**
LaBSE	1E-04
Geotrend-BERT	1E-04
mBERT	1E-05
UKR-RoBERTa Base	1E-05
XLM-RoBERTa Base	1E-05
XLM-RoBERTa Large	1E-05
Twitter-XLM-RoBERTa	1E-04
Glot500 Base	1E-06
Multilingual-E5 Large	1E-05
Multilingual-E5 Base	1E-05

Table 5: The best learning rate for the Transformer-encoder based models fine-tuned on original Ukrainian data.

G Confusion Matrices

Here, in addition to the main results, we also report the confusion matrices for the top performing models.

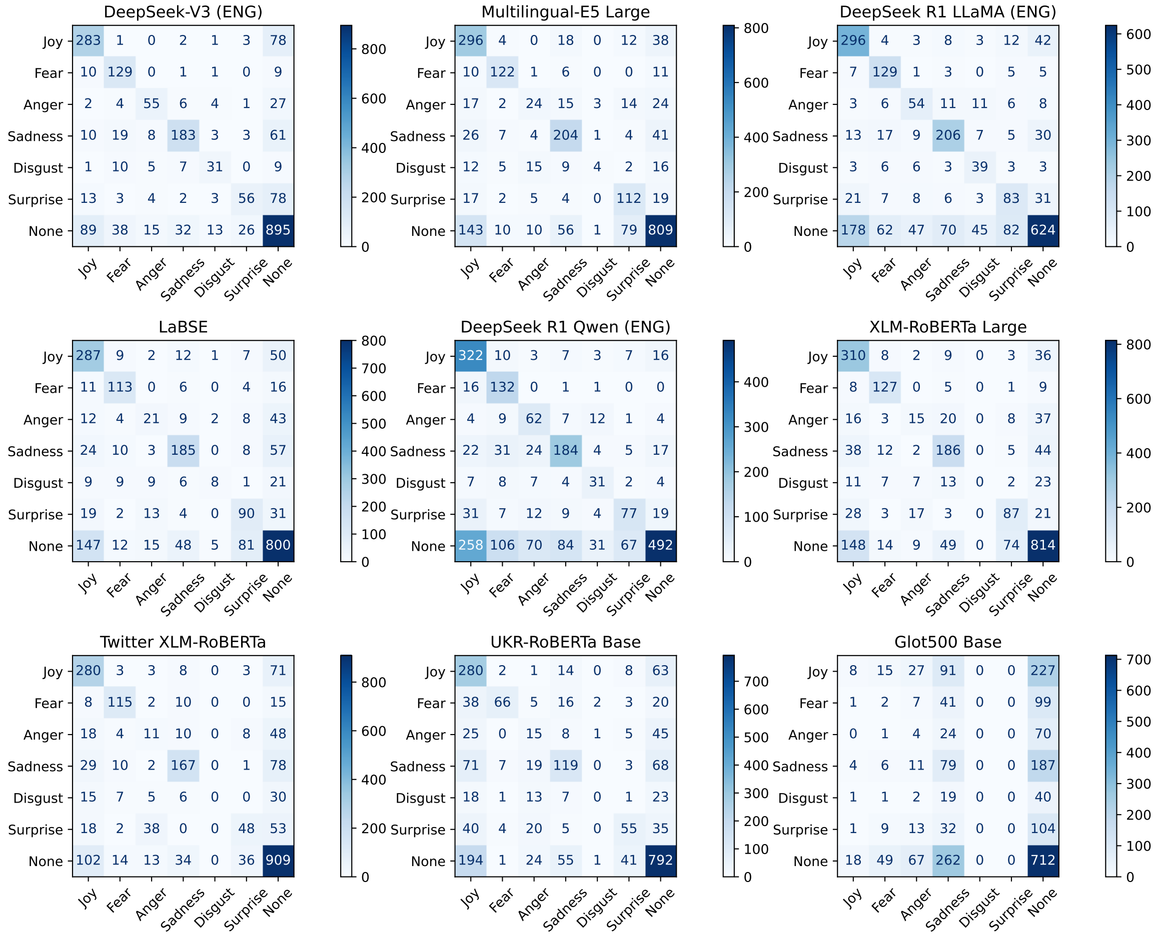


Figure 6: Confusion matrices of the top performing models fine-tuned on the EMOBENCH-UA training data.