UAReviews: A Multi-Task Ukrainian Dataset for Emotion and Intent Classification

Roman Kyslyi **Ihor Pysmennyi** **Denys Mykhailov**
Kyiv School of Economics, Ukraine
(rkyslyi,ipysmennyi,dmykhailov)@kse.org.ua

Abstract

We introduce UAReviews, a multi-task Ukrainian-language dataset for emotion and intent classification comprising 11,580 annotated texts. The dataset combines two sources: citizen reviews of government digital services provided by the Ministry of Digital Transformation of Ukraine and Ukrainian-language Telegram posts drawn from the COSMUS corpus. Each text is annotated with both an emotion label following the Ekman taxonomy (seven classes) and an intent label (five classes), making it the first publicly available Ukrainian resource for joint emotion and intent analysis. Annotation was performed by students at the Kyiv School of Economics, with a gold standard subset (20%) validated by three independent annotators achieving Krippendorff's $\alpha = 0.93$. We establish baselines using single-task and multi-task fine-tuned XLM-RoBERTa models and analyze emotion to intent correlation. Both the dataset and the baseline models are publicly available.[1]

1 Introduction

Emotion and intent detection in text are fundamental tasks in natural language processing (NLP) with wide-ranging applications, from customer feedback analysis to conversational AI systems. While substantial resources exist for high-resource languages such as English (Demszky et al., 2020; Mohammad et al., 2018), many languages still lack task-specific benchmarks. Ukrainian is no longer considered a low-resource language: thanks to growing community efforts, pre-trained models and general-purpose corpora are now available (Romanyshyn, 2023). However, the Ukrainian benchmark landscape for affective NLP is still sparse: the only existing emotion benchmark, EmoBench-UA (Dementieva et al., 2025), is multi-label and limited to social-media-style text, and to our knowledge no publicly available intent classification benchmark exists for Ukrainian. UAReviews complements EmoBench-UA by providing single-label emotion annotations together with intent labels on a different domain (government service reviews and Telegram posts).

The need for Ukrainian NLP resources has become particularly pressing in recent years. Government digitalization efforts in Ukraine, led by the Ministry of Digital Transformation, have generated large volumes of citizen feedback in Ukrainian. Understanding the emotions and intents expressed in this feedback is crucial for improving public services, yet no dedicated datasets or models existed for this purpose.

In this paper, we present UAReviews, a Ukrainian-language dataset annotated for both emotion and intent classification. Our dataset combines two complementary data sources: (1) citizen reviews of government digital services (e.g., the Diia platform) provided by the Ministry of Digital Transformation of Ukraine, and (2) Ukrainian-language posts from the COSMUS corpus (Shynkarov et al., 2025), a collection of Telegram channel messages. The inclusion of the COSMUS data is intended to add domain diversity and broaden linguistic register beyond formal government service reviews; we do not claim it balances the domain distribution, as the Telegram subset is much smaller than the government review subset (Section 3).

Our contributions are as follows:

- We release UAReviews, a publicly available multi-task Ukrainian dataset of 11,580 texts annotated for both emotion and intent classification, combining government service reviews and social media posts, with a standard train/dev/test split.
- We describe a rigorous annotation protocol

[1]Dataset and models: https://huggingface.co/KSE-RESEARCH-Group

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 12–23

using Argilla (Vila-Suero and Aranda, 2023) tool hosted on Hugging Face Spaces, with quality control through a gold standard subset achieving Krippendorff's $\alpha = 0.93$ (Krippendorff, 2011).

- We provide comprehensive baselines, including single-task and multi-task models, with per-class evaluation metrics, confusion matrices, and multi-seed variance analysis.

- We quantify the correlation between emotion and intent labels using Cramér's V and normalized mutual information, providing empirical support for joint modeling.

2 Related Work

Emotion Detection Datasets. Early work in emotion detection introduced datasets based on news headlines (Strapparava and Mihalcea, 2007) and social media (Mohammad et al., 2018). Demszky et al. (2020) released GoEmotions, a large-scale English dataset with 27 emotion categories derived from Reddit comments. These resources have driven progress in emotion detection for English, but comparable resources for most other languages remain scarce.

Emotion taxonomies in NLP typically follow either Ekman (1992), who proposed six basic emotions (anger, disgust, fear, happiness, sadness, surprise), or Plutchik (1980), who extended this to eight primary emotions organized in a wheel. Our annotation scheme follows the Ekman taxonomy, adopting the six basic emotions supplemented with a *Neutral* category.

For Ukrainian specifically, EmoBench-UA (Dementieva et al., 2025) is the only other emotion detection benchmark. It was created using crowdsourcing on Toloka and formulated as a multi-label task, with the best-performing model (DeepSeek-V3) reaching a macro F1 of 0.65. UAReviews differs in two key aspects: (i) it targets single-label classification, reflecting the dominant emotion per text, and (ii) it additionally provides intent annotations, enabling joint emotion-intent research.

Intent Classification. Intent classification has been studied primarily in the context of task-oriented dialogue systems (Casanueva et al., 2020). While intent detection datasets exist for English and a few other languages, to our knowledge no publicly available intent classification dataset exists for Ukrainian.

Ukrainian NLP Resources. Ukrainian NLP has seen growing attention in recent years (Romanyshyn, 2023), and the language can no longer be considered low-resource in terms of pre-trained models. Prior work has addressed sentiment analysis, with Shynkarov et al. (2025) introducing the COSMUS corpus—a 12,224-text collection from Telegram channels and product-review sites—for Ukrainian social media sentiment analysis in code-switching contexts. EmoBench-UA (Dementieva et al., 2025) provides multi-label emotion annotations for Ukrainian. UAReviews differs in its focus on single-label classification, the addition of intent annotations, and a different domain (government service reviews), making the two datasets complementary rather than redundant.

3 Dataset

3.1 Data Sources

The UAReviews dataset is constructed from two complementary sources:

Government Service Reviews. The primary data source consists of citizen reviews of government digital services, provided by the Ministry of Digital Transformation of Ukraine. These reviews cover a range of public services delivered through digital platforms (e.g., Diia) and include feedback on service quality, usability, and citizen satisfaction. The texts are exclusively in Ukrainian and represent semi-formal register. Each review is accompanied by a numerical rating.

COSMUS Telegram Posts. To increase domain diversity and balance the dataset, we incorporate Ukrainian-language posts from the COSMUS dataset (Shynkarov et al., 2025). COSMUS is a corpus of Telegram channel messages originally compiled for sentiment analysis research, containing texts in Ukrainian, Russian, and code-switched varieties. We selected only the Ukrainian-language subset of COSMUS. These texts represent a more informal, social-media register and cover a broader range of topics than the government service or institutions reviews.

3.2 Annotation Scheme

Each text in UAReviews is annotated along two dimensions:

Emotion Labels. We adopt an emotion taxonomy grounded in the established psychological framework of Ekman (1992). Each text is assigned one of seven emotion labels: *Happiness*, *Anger*, *Sadness*, *Fear*, *Surprise*, *Disgust*, and *Neutral* (no dominant emotion). The label set was chosen to balance granularity with practical annotator reliability. The *Neutral* category captures texts that express factual content or mixed emotions without a clearly dominant affective signal.

Intent Labels. Each text is additionally annotated with an intent label (stored as `final_category` in the dataset) capturing the communicative purpose of the text. Intent categories were designed to reflect the pragmatic functions common in citizen feedback and social media discourse. The five intent categories are: *Gratitude / Positive Feedback*, *Complaint / Dissatisfaction*, *Question / Request for Help*, *Neutral Comment*, and *Suggestion / Idea*.

3.3 Annotation Process

Annotation was carried out by students at the Kyiv School of Economics using Argilla (Vila-Suero and Aranda, 2023), an open-source data annotation platform hosted on Hugging Face Spaces.

We report inter-annotator agreement using Krippendorff's α (Krippendorff, 2011), a chance-corrected reliability measure suitable for multiple annotators and nominal data (Artstein and Poesio, 2008). Because the dataset has two annotation dimensions, we compute α separately for emotion and for intent and report the average of the two; the values are tightly clustered around the reported figure for both dimensions.

Gold Standard Set. Approximately 20% of the dataset (∼2,316 texts) was designated as the gold standard set. Each text in this subset was independently annotated by three human annotators. The inter-annotator agreement on the gold set was Krippendorff's $\alpha = 0.93$, indicating near-perfect agreement and high annotation quality.

Main Annotation. The remaining 80% of the dataset (∼9,264 texts) was annotated by two human annotators with assistance from Gemini (Gemini Team, 2023), a large language model used as a third annotator. The inter-annotator agreement on this portion, including the LLM-assisted annotations, was $\alpha = 0.87$. Final labels were determined by majority vote across the three annotations (two human + one LLM). Crucially, the gold standard set (20%, fully human-annotated with $\alpha = 0.93$) serves as a quality control benchmark: the small gap between $\alpha = 0.93$ (gold) and $\alpha = 0.87$ (LLM-assisted) provides evidence that the LLM-assisted portion does not substantially degrade annotation quality. The gold set also enables future work to directly compare model performance on fully human-annotated vs. LLM-assisted subsets, should per-annotator labels be released.

Statistic	Value
Total texts	11,580
Gold standard subset (20%)	∼2,316
Main subset (80%)	∼9,264
Language	Ukrainian
Annotation dimensions	emotion, intent
Annotators (gold set)	3 humans
Annotators (main set)	2 humans + LLM
Krippendorff's α (gold set)	0.93
Krippendorff's α (main set)	0.87

Table 1: Overview of the UAReviews dataset.

Discussion of Agreement Levels. We acknowledge that $\alpha = 0.93$ is unusually high for subjective emotion annotation. We attribute this to several factors: (1) the review domain often produces texts with clear emotional signals (e.g., explicit gratitude or complaints), (2) the seven-class taxonomy based on Ekman's well-established framework is relatively coarse-grained, and (3) the annotator training and clear guidelines (Appendix D) promoted consistent interpretations. The dominant classes (*Happiness* at 65.3%, *Gratitude / Positive Feedback* at 64.2%) are typically unambiguous, which likely contributes to the high overall agreement.

3.4 Dataset Statistics

Table 1 summarizes the key statistics of the UAReviews dataset.

Table 2 shows the emotion label distribution. The dataset exhibits a pronounced class imbalance: *Happiness* dominates at 65.3%, and *Gratitude / Positive Feedback* at 64.2%. We emphasize that this distribution is not an artifact of sampling bias - it reflects the real-world distribution of citizen feedback on government digital services. Satisfied users naturally leave positive reviews more

Emotion	Count	%
Happiness	7,557	65.3
Anger	2,264	19.5
Neutral	1,117	9.6
Sadness	424	3.7
Disgust	106	0.9
Surprise	57	0.5
Fear	55	0.5
Total	**11,580**	**100.0**

Table 2: Emotion label distribution in UAReviews, sorted by frequency.

Intent	Count	%
Gratitude / Positive Feedback	7,440	64.2
Complaint / Dissatisfaction	2,730	23.6
Question / Request for Help	615	5.3
Neutral Comment	418	3.6
Suggestion / Idea	377	3.3
Total	**11,580**	**100.0**

Table 3: Intent label distribution in UAReviews, sorted by frequency.

frequently, a well-documented phenomenon in review and customer feedback datasets (Hu and Liu, 2004; McAuley and Leskovec, 2013). Similar positive skews appear in app store reviews (where 5-star ratings routinely exceed 60%), customer satisfaction surveys, and product feedback platforms. Artificially rebalancing the dataset (e.g., by oversampling negative reviews or discarding positive ones) would misrepresent the true distribution that deployed systems must handle.

The inclusion of COSMUS Telegram posts was motivated by the need to diversify the emotional and intent distribution: the COSMUS subset contains 67.1% *Neutral* and 18.6% *Sadness* (compared to 9.3% and 3.6% respectively in government reviews), providing a complementary signal at the per-class level. We caution, however, that with only 170 COSMUS texts versus 11,410 government reviews, this addition adds register and topical diversity rather than meaningfully balancing the overall domain distribution. Expanding the COSMUS component is a priority for future dataset releases and would also enable proper cross-domain evaluation between the two sources. The tail classes - *Sadness* (3.7%), *Disgust* (0.9%), *Surprise* (0.5%), and *Fear* (0.5%) - are substantially underrepresented, posing challenges for classification. To mitigate this, we employ class-weighted cross-entropy loss (Section 4) and report both macro and weighted F1 to make the impact of imbalance transparent.

Figure 1 visualizes the class distributions for both annotation dimensions.

Each record in the released dataset contains the following fields: a unique identifier (`id`), a numerical `rating` (where applicable), the text `content`, a `source` indicator (government reviews or COSMUS), the `final_emotion` label, the `final_category` (intent) label, the text `length`, and the `split` assignment (train, dev, or test).

3.5 Emotion-Intent Correlation

To quantify the relationship between emotion and intent labels, we compute Cramér's V and normalized mutual information (NMI) on the full dataset. The cross-tabulation (Figure 2, Appendix E) reveals strong alignment: *Happiness* maps almost exclusively to *Gratitude / Positive Feedback*, while *Anger* aligns with *Complaint / Dissatisfaction*. The overall association is strong (Cramér's $V = \mathbf{0.63}$, bias-corrected $V = 0.63$; NMI $= 0.70$; $\chi^2 = 18{,}603.84$, $p < 0.001$), confirming that the two annotation dimensions are statistically dependent but not redundant - tail emotion classes such as *Fear*, *Surprise*, and *Disgust* distribute across multiple intent categories.

4 Experiments

4.1 Data Splits

We partition the dataset into *train* (70%; 8,106 texts), *dev* (15%; 1,737 texts), and *test* (15%; 1,737 texts) using stratified sampling based on the joint emotion-intent label. All hyperparameter and model selection decisions were made on the dev set; the held-out test set was used only for the final evaluations reported in this section. We report mean and standard deviation across five random seeds on this same held-out test set; no model selection was performed on test data. The split preserves class proportions across all three subsets (see Appendix B for per-class breakdowns).

4.2 Model

For all experiments, we use *XLM-RoBERTa-base-uk*, a Ukrainian-adapted variant of XLM-

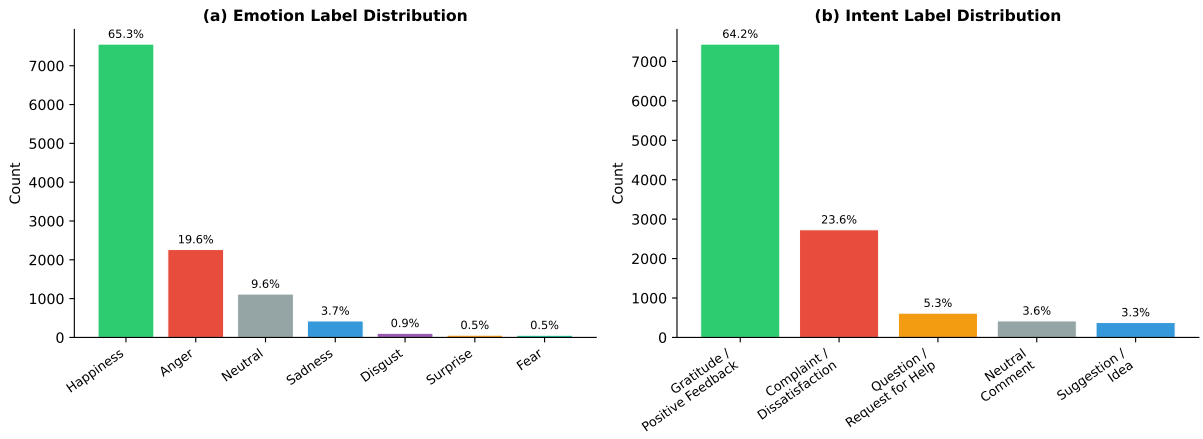

Figure 1: Class distribution for (a) emotion and (b) intent labels in UAReviews. Both dimensions exhibit a long-tail distribution dominated by positive classes, which is characteristic of review and feedback data.

RoBERTa (Conneau et al., 2020). Unlike the original multilingual model (which contains 470M parameters with embeddings for 100 languages), this checkpoint retains only Ukrainian and English vocabulary embeddings, reducing the model to 110M parameters while preserving the encoder's representational capacity for Ukrainian.

4.3 Single-Task Baselines

We train separate models for emotion classification (7 classes) and intent classification (5 classes). A standard `XLMRobertaForSequenceClassification` head is added on top of the pre-trained encoder. Both models share the same hyperparameter configuration (Table 4), fine-tuned using PyTorch Lightning with the Hugging Face Transformers library (Wolf et al., 2020).

We employ separate learning rates for the classification head (5e-5) and the pre-trained embeddings (9e-5), and class-weighted cross-entropy loss to mitigate class imbalance. Weights are computed as inverse frequency raised to a power of 0.2, then normalized to sum to 1.

4.4 Multi-Task Baseline

To test whether joint modeling improves over single-task training, we implement a shared-encoder multi-task model: a single XLM-RoBERTa encoder feeds into two separate classification heads (one for emotion, one for intent), each consisting of a dense layer with tanh activation followed by a linear projection. The total loss is an equally weighted sum of the two task losses ($\lambda_{\text{emo}} = \lambda_{\text{int}} = 0.5$). We report the average macro F1 across both tasks as the primary metric.

Hyperparameter	Value
Pre-trained model	xlm-roberta-base-uk
Classifier LR	5e-5
Embedding LR	9e-5
Optimizer	AdamW
Weight decay	0.005
Effective batch size	128
Max epochs	10
Early stopping patience	5 (on dev F1)
Scheduler	Cosine with warmup
Warmup ratio	0.1
Precision	16-bit mixed
Gradient clipping	1.0 (norm)
Class weight power	0.2

Table 4: Shared hyperparameters for both tasks.

4.5 Results

Table 5 presents the main results. We report macro F1 (mean $\pm$ standard deviation across 5 random seeds) on the held-out test set; per-seed numbers are provided in Appendix A. The single-task models achieve strong performance, with intent classification (macro F1 = 0.93 $\pm$ 0.05) outperforming emotion classification (0.81 $\pm$ 0.15). The high variance on emotion reflects the difficulty of rare-class prediction: across the five seeds, emotion macro F1 ranges from 0.51 (worst seed) to 0.92 (best seed), with the worst seed driven down by tail-class collapse. The per-class results in Tables 6 and 7 are reported for seed 42, which produced near-mean performance (macro F1 $\approx$ 0.82

Model	Emo $F1_M$	Int $F1_M$
Single-task	**0.81 ± 0.15**	**0.93 ± 0.05**
Multi-task (shared)	0.55 ± –	0.81 ± –

Table 5: Test-set macro F1 (mean ± std over 5 seeds) for emotion and intent classification. $F1_M$ denotes macro-averaged F1 (Opitz and Burst, 2021).

Emotion	P	R	F1
Happiness	98.1	99.0	98.6
Anger	96.6	94.9	95.8
Neutral	88.7	91.1	89.9
Sadness	83.1	81.2	82.1
Disgust	72.7	76.2	74.4
Surprise	100.0	63.6	77.8
Fear	100.0	36.4	53.3
Macro avg	91.3	77.5	81.7
Weighted avg	96.1	96.1	96.0

Table 6: Per-class precision (P), recall (R), and F1 for **emotion** classification on the test set (single-task model, seed 42, used here as a representative near-mean run).

on emotion and 0.81 on intent); we use this seed as a representative single-run breakdown rather than as the worst- or best-case run. The multi-task model underperforms single-task baselines, particularly for emotion (0.55 vs. 0.81 macro F1), suggesting that the shared encoder may not effectively balance both objectives with equal loss weighting. This is an area for future work.

Tables 6 and 7 present per-class precision, recall, and F1 for both tasks. As expected, the high-frequency classes (*Happiness*, *Gratitude / Positive Feedback*) achieve the strongest F1 scores, while the tail classes (*Fear*, *Surprise*, *Disgust*) show lower performance. Fear achieves the lowest F1 (53.3) with perfect precision but only 36.4% recall, indicating the model identifies Fear when it predicts it but misses most instances. Surprise similarly suffers from low recall (63.6%). The confusion matrix confirms that tail-class errors are not simply majority-class predictions: Fear is most often confused with Sadness and Neutral, while Disgust errors distribute across Anger and Fear.

Figures 3 and 4 (Appendix E) show the confusion matrices. The emotion confusion matrix reveals that errors are concentrated among tail classes, which tend to be confused with the dominant *Happiness* class. The intent confusion matrix shows a cleaner diagonal, consistent with the higher macro F1 for this task.

Intent	P	R	F1
Gratitude / Pos. Fb.	96.3	97.6	96.9
Complaint / Dissat.	95.0	89.6	92.2
Question / Help	86.7	90.2	88.4
Neutral Comment	57.0	63.1	59.9
Suggestion / Idea	70.8	68.0	69.4
Macro avg	81.2	81.7	81.4
Weighted avg	93.2	93.1	93.1

Table 7: Per-class precision (P), recall (R), and F1 for **intent** classification on the test set (single-task model, seed 42, used here as a representative near-mean run). Note that the macro F1 of 81.4 on this individual seed is consistent with the 5-seed mean of 0.93 reported in Table 5: most seeds reach ∼0.96 macro F1, with seed 42 being a lower-end run.

5 Analysis

Minority Class Performance. The per-class results (Tables 6 and 7) allow direct assessment of whether the model's aggregate F1 is driven primarily by head classes. While the gap between macro and weighted F1 quantifies this effect, the tail-class F1 scores are the critical metric. Despite extreme imbalance (e.g., *Fear*: 55 examples, 0.5%), the class-weighted loss with power 0.2 provides meaningful correction.

Class Weight Ablation. We chose a class weight power of 0.2 based on preliminary experiments. A power of 0 (uniform weighting) leads to majority-class bias; powers above 0.5 can overfit to rare classes. We acknowledge that we do not present a systematic ablation over this hyperparameter; future work could compare with focal loss (Lin et al., 2017) or resampling strategies.

Multi-Task Performance Collapse. The multi-task model exhibits a severe performance drop on emotion classification (macro F1: 0.81 → 0.55), while intent performance remains comparable to the single-task baseline (0.81 vs. 0.93). This asymmetric degradation - where one task collapses while the other is preserved - needs careful analysis. We identify several plausible contributing factors:

(1) Task difficulty asymmetry and head dominance. The intent task (5 classes, less imbalanced,

macro F1 = 0.93 single-task) is substantially easier than the emotion task (7 classes, extreme imbalance, macro F1 = 0.81). With equal loss weighting ($\lambda = 0.5$), the shared encoder may converge toward representations that favor the easier intent task, as intent gradients dominate early training and shape the shared representation space before the harder emotion task can establish useful features. This is a well-known failure mode in multi-task learning (Chen et al., 2018).

(2) Strong label correlation as a double-edged sword. The high correlation between emotion and intent (Cramér's $V = 0.63$, NMI $= 0.70$) means that intent labels are partially predictive of emotion. Rather than providing complementary signal, the shared encoder may learn to rely on intent-discriminative features that are sufficient for the coarse emotion–intent mapping (e.g., *Happiness* $\leftrightarrow$ *Gratitude*) but insufficient for distinguishing tail emotion classes that span multiple intents (e.g., *Fear* appears in both *Complaint* and *Question*).

(3) Representation conflict on tail classes. The emotion tail classes (*Fear*, *Surprise*, *Disgust*: collectively 1.9% of data) require fine-grained distinctions that the shared encoder may sacrifice to improve performance on the majority classes shared across both tasks. The per-class results confirm this: multi-task emotion F1 for *Disgust* drops to 22.2 (from 74.4 single-task) and *Fear* to 30.0 (from 53.3), while *Happiness* remains high at 95.9.

(4) Loss imbalance. Although both task losses use class weights, the equal $\lambda = 0.5$ weighting does not account for the difference in task difficulty or convergence rate. The intent head converges faster and continues to dominate gradient updates throughout training.

We did not ablate over λ values, encoder freezing schedules, or hierarchical architectures (e.g., predicting intent first and conditioning emotion on intent) in this work. These directions represent important areas for future work.

We present the multi-task result as a baseline, and we note that its performance itself is informative: it demonstrates that naive shared-encoder multi-task learning is insufficient for this dataset and that the strong emotion-intent correlation does not guarantee multi-task gains.

Annotation Quality: Per-Class Considerations. While the overall $\alpha = 0.93$ on the gold set indicates high agreement, this aggregate metric may mask lower agreement on inherently subjective categories. Emotions like *Neutral* (which serves as a residual category) and *Sadness* (which can overlap with *Anger* in complaint contexts) are likely harder to annotate consistently. Future releases of UAReviews will include per-class α values. We also note that the high proportion of unambiguous positive reviews (65.3% *Happiness*) contributes to the high aggregate agreement.

6 Discussion

UAReviews addresses a significant gap in Ukrainian NLP resources by providing the first publicly available dataset for joint emotion and intent classification. Several design decisions merit further discussion.

Hybrid Annotation Strategy. Our use of an LLM (Gemini) as a third annotator alongside two human annotators for 80% of the dataset is a design choice to balance annotation cost with quality. The key safeguard is the gold standard set: 20% of the dataset was annotated entirely by three human annotators ($\alpha = 0.93$), providing an LLM-free reference point. The marginal drop to $\alpha = 0.87$ on the LLM-assisted portion suggests that Gemini's annotations are largely consistent with human judgments in this domain. Importantly, because final labels are determined by majority vote (two humans + one LLM), Gemini can only influence the label when the two human annotators disagree - it cannot override a human consensus. This architecture means the LLM acts as a tiebreaker rather than a primary annotator. We acknowledge that without per-annotator labels, we cannot quantify potential LLM biases (e.g., toward majority classes) and plan to release per-annotator labels in a future version to enable such analysis.

Ukrainian-Adapted Pre-training. Our choice of `ukr-models/xlm-roberta-base-uk` - a vocabulary-trimmed variant of XLM-RoBERTa that retains only Ukrainian and English embeddings - over the full multilingual checkpoint reflects a practical trade-off: the trimmed model is significantly smaller (110M vs. 470M parameters) while preserving the encoder's capacity for Ukrainian.

Joint Modeling Potential. The strong statistical dependence between emotion and intent (Section 3.5) motivates multi-task learning, yet our

shared-encoder baseline demonstrates that naive approaches fail (Section 5). The performance collapse on emotion suggests that effective joint modeling requires architectural innovations: task-specific loss weighting ($\lambda_{\text{emo}} > \lambda_{\text{int}}$) to compensate for difficulty asymmetry, gradient balancing methods (Chen et al., 2018; Kendall et al., 2018), hierarchical prediction (e.g., predicting intent first and conditioning emotion on intent), or cross-attention mechanisms between task heads. The dataset's dual annotations make it a valuable testbed for such research.

7 Conclusion

We have presented UAReviews, a multi-task Ukrainian-language dataset of 11,580 texts annotated for emotion classification (7 classes) and intent classification (5 classes). The dataset combines citizen reviews of government services with Ukrainian Telegram posts from the COSMUS corpus, annotated with a rigorous quality control protocol achieving Krippendorff's $\alpha = 0.93$ on the gold standard subset. We provide comprehensive baselines - single-task and multi-task, with per-class metrics, confusion matrices, and multi-seed variance analysis, including a detailed analysis of multi-task performance collapse. Both the dataset and the models are publicly released to support Ukrainian NLP research.

UAReviews fills a gap in Ukrainian NLP by providing the first jointly annotated emotion–intent dataset, enabling research at the intersection of affective computing and pragmatic analysis for an underserved language. Our experiments reveal two key findings. First, single-task fine-tuning of a vocabulary-trimmed XLM-RoBERTa achieves strong intent classification (macro F1 = 0.93) but exposes the difficulty of tail-class emotion detection under extreme imbalance, where seed sensitivity alone can swing macro F1 by 0.4. Second, the failure of naive shared-encoder multi-task learning—despite strong statistical dependence between labels—demonstrates that label correlation is a necessary but insufficient condition for multi-task gains.

Limitations

The UAReviews dataset has several limitations. First, the dataset exhibits significant class imbalance, which affects tail-class performance. Although class-weighted loss partially addresses this, we do not present ablations comparing it with alternatives such as focal loss or oversampling. Second, the dataset is drawn from only two domains (government service reviews and Telegram posts), with the Telegram subset (170 texts) being too small for reliable cross-domain evaluation. Expanding this component would improve the dataset's utility for domain adaptation research. Third, the high α values may partly reflect the dominance of unambiguous positive reviews rather than genuine ease of annotation across all categories. Fourth, we report results from a single model architecture, exploring larger models, ensembles, or additional pre-training may be beneficial.

Ethics Statement

The government service reviews used in this dataset were provided by the Ministry of Digital Transformation of Ukraine for research purposes. All texts were anonymized prior to annotation to remove personally identifiable information. The Telegram posts were sourced from the publicly available COSMUS dataset (Shynkarov et al., 2025). Annotators were students at the Kyiv School of Economics who participated voluntarily under a formal arrangement with the institution; the annotation activity was not unpaid labor. The annotation process was designed to avoid exposing annotators to harmful content. We release the dataset under a CC-BY-4.0 licence.

Acknowledgements

We thank the Ministry of Digital Transformation of Ukraine for providing access to the citizen review data, and the student annotators at the Kyiv School of Economics for their careful annotation work. We also thank the anonymous reviewers of UNLP 2026 for their constructive feedback, which helped improve this paper.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 794–803.

Alexis Conneau, Kartik 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.

Daryna Dementieva, Nikolay Babakov, and Alexander Fraser. 2025. EmoBench-UA: A benchmark dataset for emotion detection in Ukrainian. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2025–2048.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172.

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.

Jürgen Opitz and Sebastian Burst. 2021. Macro F1 and macro F1. *arXiv preprint arXiv:1911.03347*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Mariana Romanyshyn. 2023. Proceedings of the second Ukrainian natural language processing workshop (UNLP). In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*.

Yurii Shynkarov, Veronika Solopova, and Vera Schmitt. 2025. Improving sentiment analysis for Ukrainian social media code-switching data. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 179–193.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Daniel Vila-Suero and Francisco Aranda. 2023. Argilla: Open-source framework for data-centric NLP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moe, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacin Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

A Training Details

All models were trained using PyTorch Lightning with class-weighted cross-entropy loss. Class weights were computed from inverse frequency weighting (power = 0.2, normalized). Tables 8 and 9 list the weights for each task.

Emotion Class	Weight
Happiness	0.076
Anger	0.096
Neutral	0.111
Sadness	0.135
Disgust	0.178
Surprise	0.201
Fear	0.203

Table 8: Class weights for emotion classification.

Intent Class	Weight
Gratitude / Positive Feedback	0.134
Complaint / Dissatisfaction	0.164
Question / Request for Help	0.221
Neutral Comment	0.238
Suggestion / Idea	0.243

Table 9: Class weights for intent classification.

The training split preserves the original class distribution via stratified sampling. All models were trained for a maximum of 10 epochs with early stopping (patience = 5) based on dev macro F1 (Opitz and Burst, 2021). Results are reported on the held-out test set. Multi-seed experiments use seeds $\{42, 123, 456, 789, 2024\}$.

B Data Splits

The dataset is split into train (70%), dev (15%), and test (15%) via stratified sampling on the joint emotion–intent label. Table 10 shows the sizes.

	Train	Dev	Test
Texts	8,106	1,737	1,737
%	70.0	15.0	15.0

Table 10: Dataset split sizes.

C Data Fields

Each record in the UAReviews dataset contains the following fields:

- `id`: Unique identifier for the text.
- `rating`: Numerical rating (applicable to government service reviews).
- `content`: The raw text of the review or post.
- `source`: Indicator of the data source (government reviews or COSMUS).
- `final_emotion`: The consensus emotion label (one of seven classes).
- `final_category`: The consensus intent/category label.
- `length`: The character length of the text.
- `split`: Train/dev/test split assignment.

D Annotation Guidelines

The following guidelines were provided to annotators via the Argilla platform. Annotators were instructed to read each text carefully and assign exactly one emotion label and one intent label. When a text conveyed multiple emotions or intents, annotators were asked to select the dominant one - the emotion or intent that most strongly characterized the text as a whole. In cases of genuine ambiguity, annotators were encouraged to use the *Neutral* emotion label or the *Neutral Comment* intent label.

D.1 Emotion Label Definitions

The emotion taxonomy follows Ekman (1992), with the addition of a *Neutral* category. Definitions and representative examples are provided below in the original Ukrainian with English translations.

Happiness. The text expresses positive feelings such as joy, satisfaction, gratitude, delight, or contentment. The author conveys a favorable experience or outcome.

Example: « Все працює ідеально, дякую! Отримав документ за 5 хвилин без жодної черги. » ("Everything works perfectly, thank you! Got my document in 5 minutes without any queue.")

Anger. The text expresses irritation, frustration, outrage, or hostility. The author is displeased and may direct negativity at a service, person, or situation.

Example: « Жахливе місце. Жахлива черга. Ніхто нічого підказати не може. » ("Terrible place. Terrible queue. Nobody can help with anything.")

Sadness. The text expresses grief, disappointment, sorrow, or melancholy. The tone is somber or dejected rather than angry.

Example: « На жаль, я не зміг отримати потрібну послугу. Прикро, що нічого не змінилося. » ("Unfortunately, I couldn't get the service I needed. It's disappointing that things haven't improved.")

Fear. The text expresses worry, anxiety, concern, or apprehension about a potential negative outcome or threat.

Example: « Хвилююся, що мої персональні дані можуть бути в небезпеці » ("I'm worried

that my personal data might be at risk")

Surprise. The text expresses astonishment or unexpectedness, either positive or negative. The author did not anticipate the experience described.

Example: «Зовсім не очікував цього - процес був неймовірно швидким, я був у шоці!» ("I didn't expect this at all - the process was incredibly fast, I was shocked!")

Disgust. The text expresses revulsion, contempt, or strong aversion. The author finds the situation deeply unpleasant or morally objectionable.

Example: «Умови в офісі огидні - брудно, переповнено, а персонал грубий.» ("The conditions in the office were revolting - dirty, overcrowded, and the staff were rude.")

Neutral. The text does not express a clearly dominant emotion. It may be purely factual, informational, or contain mixed emotions that do not resolve to a single category.

Example: «Офіс працює з 9:00 до 18:00 у будні. Потрібно мати при собі паспорт та ідентифікаційний код.» ("The office is open from 9:00 to 18:00 on weekdays. You need to bring your passport and ID code.")

D.2 Intent Label Definitions

The intent labels capture the communicative purpose of the text. Definitions and representative examples are provided below.

Gratitude / Positive Feedback. The author's primary purpose is to express thanks, appreciation, or provide a positive evaluation of a service, product, or experience.

Example: «Дуже дякую! Працівники були дуже привітні і все зробили швидко.» ("Thank you so much! The staff were very helpful and everything was done quickly.")

Complaint / Dissatisfaction. The author's primary purpose is to express dissatisfaction, report a problem, or criticize a service, product, or experience.

Example: «Чекаю вже три години і досі не обслужили. Це неприпустимо.» ("I've been waiting for three hours and still haven't been served. This is unacceptable.")

Question / Request for Help. The author's primary purpose is to seek information, ask a question, or request assistance with a specific issue.

Example: «Підкажіть, які документи потрібно взяти для заміни посвідчення особи?» ("Can someone tell me which documents I need to bring for the ID card replacement?")

Neutral Comment. The author's primary purpose is to share factual information, make an observation, or provide a comment without a clear positive, negative, or help-seeking intent.

Example: «Нове відділення відкрилося минулого тижня на Хрещатику.» ("The new branch opened last week on Khreshchatyk Street.")

Suggestion / Idea. The author's primary purpose is to propose an improvement, offer constructive feedback, or share an idea for how a service or process could be enhanced.

Example: «Було б чудово, якби додали можливість записуватися онлайн, щоб уникнути черг.» ("It would be great if you added an option to schedule appointments online to avoid the queues.")

D.3 Special Instructions

Annotators were provided with the following additional guidance:

- **Sarcasm and irony:** If a text uses sarcasm (e.g., «Чудовий сервіс, лише чотири години почекав!» — "Great service, only had to wait four hours!"), annotate based on the *intended* meaning, not the literal surface form. In this case, the emotion would be *Anger* and the intent *Complaint / Dissatisfaction*.
- **Mixed signals:** If a text contains both positive and negative elements (e.g., «Працівники привітні, але чекати занадто довго.» — "The staff were friendly but the wait was too long"), select the emotion and intent that dominate the overall tone.
- **Short texts:** For very short texts (e.g., single words or emoji-only), annotate based on the available signal. If no clear emotion can be inferred, use *Neutral*.
- **Code-switching:** Some texts from the COSMUS subset may contain Russian words or phrases. Annotate based on the overall meaning regardless of the language used within the text.

E Additional Figures

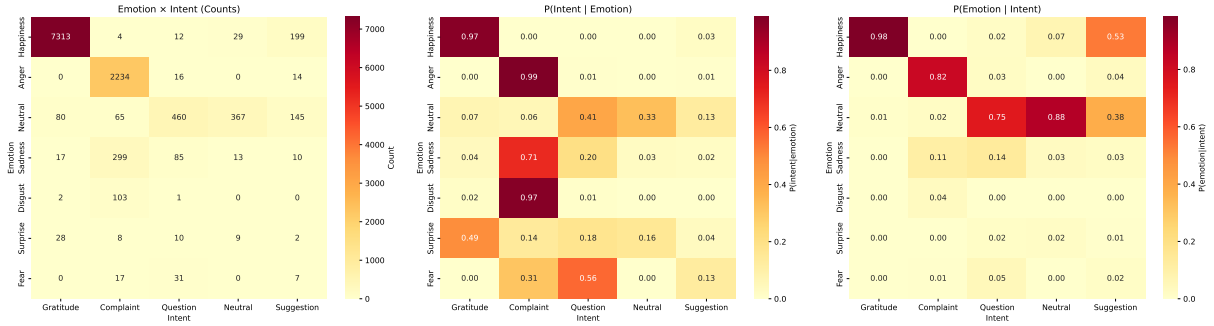

Figure 2: Emotion–intent cross-tabulation heatmap showing raw counts (left), P(intent|emotion) (center), and P(emotion|intent) (right).

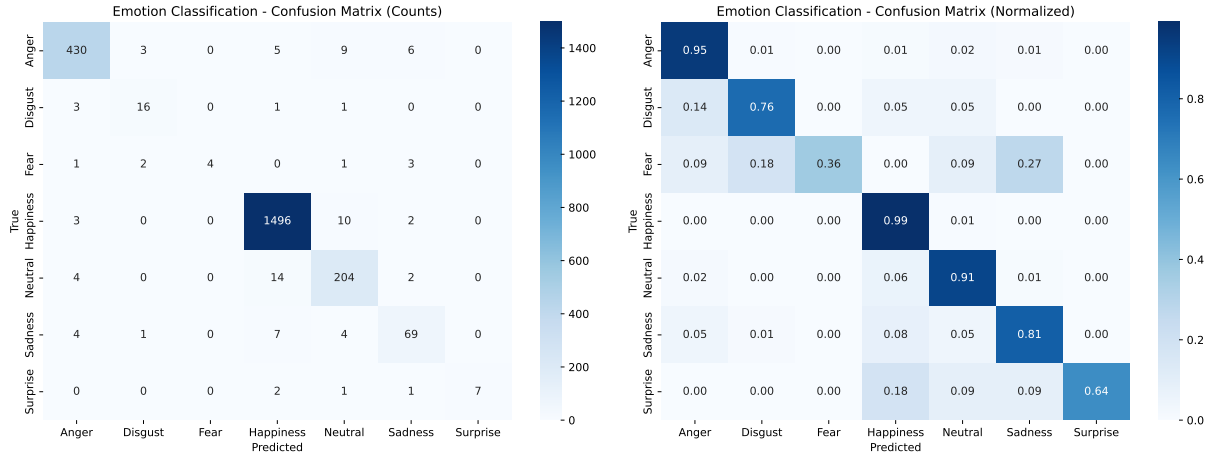

Figure 3: Confusion matrix for emotion classification (normalized by true class). The model achieves strong performance on head classes but struggles to distinguish tail classes from *Happiness* and *Neutral*.

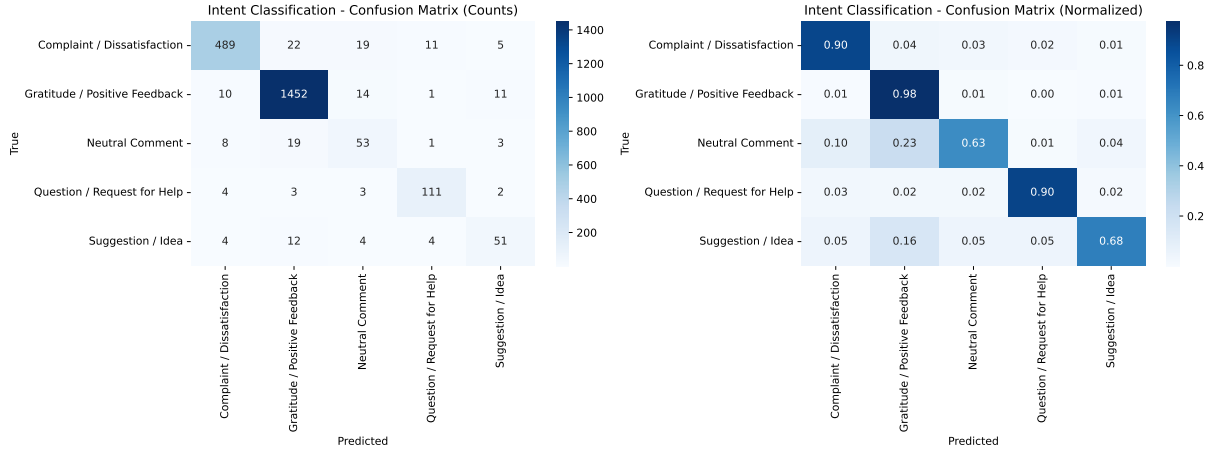

Figure 4: Confusion matrix for intent classification (normalized by true class).