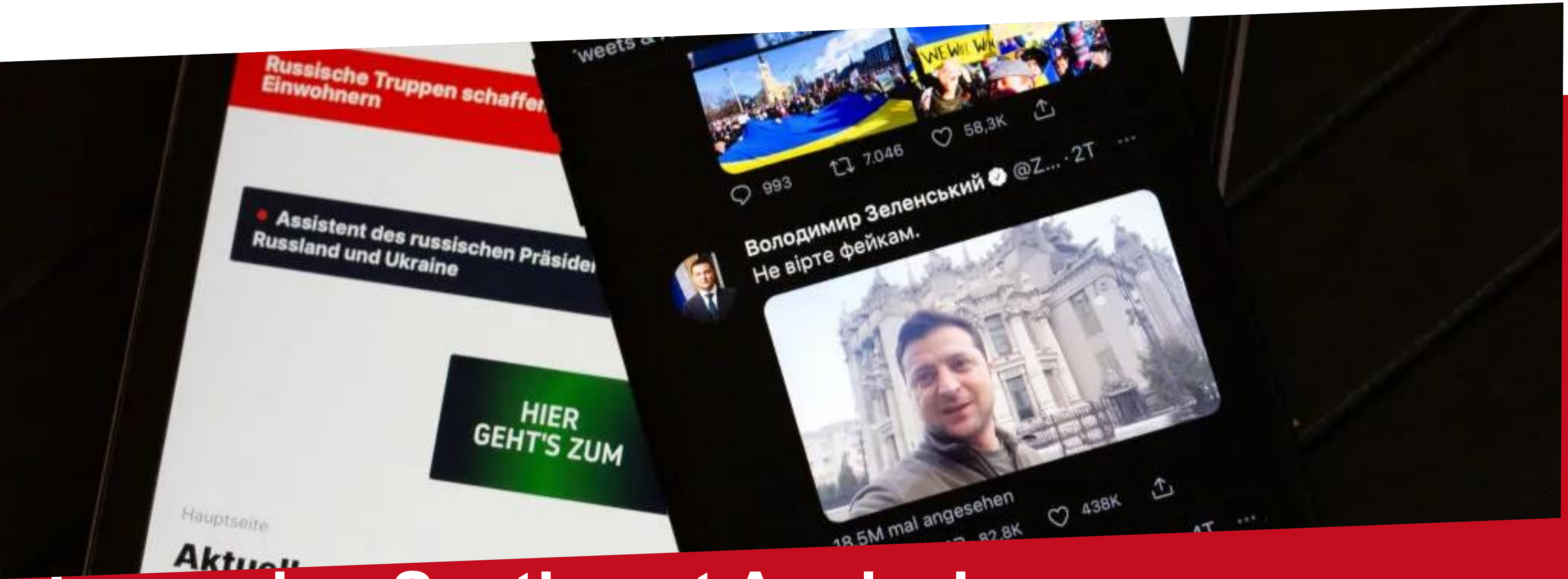


Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data | UNLP 2025

Yurii Shynkarov, Veronika Solopova, Vera Schmitt

Veronika Solopova | Quality & Usability Lab, XplaiNLP group | **UNLP Workshop 2025**

01. 08. 2025

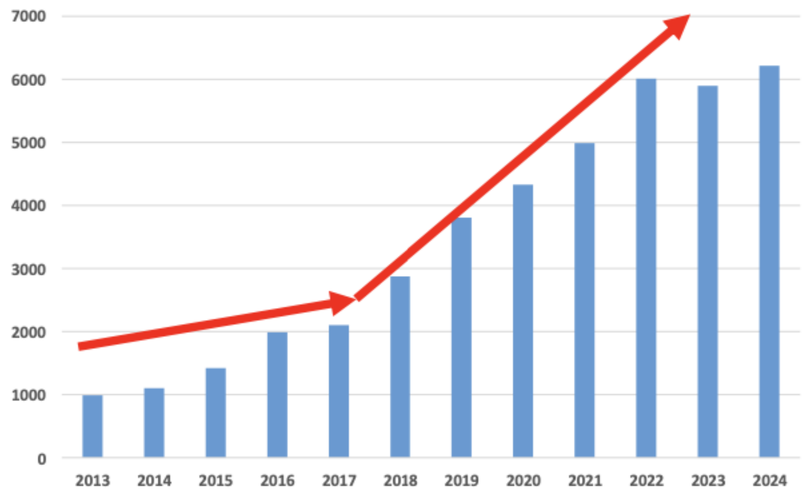
Motivation

Sentiment analysis is…

the process of computationally categorizing opinions expressed in a piece of information (**Liu, Bing.** ,2012)

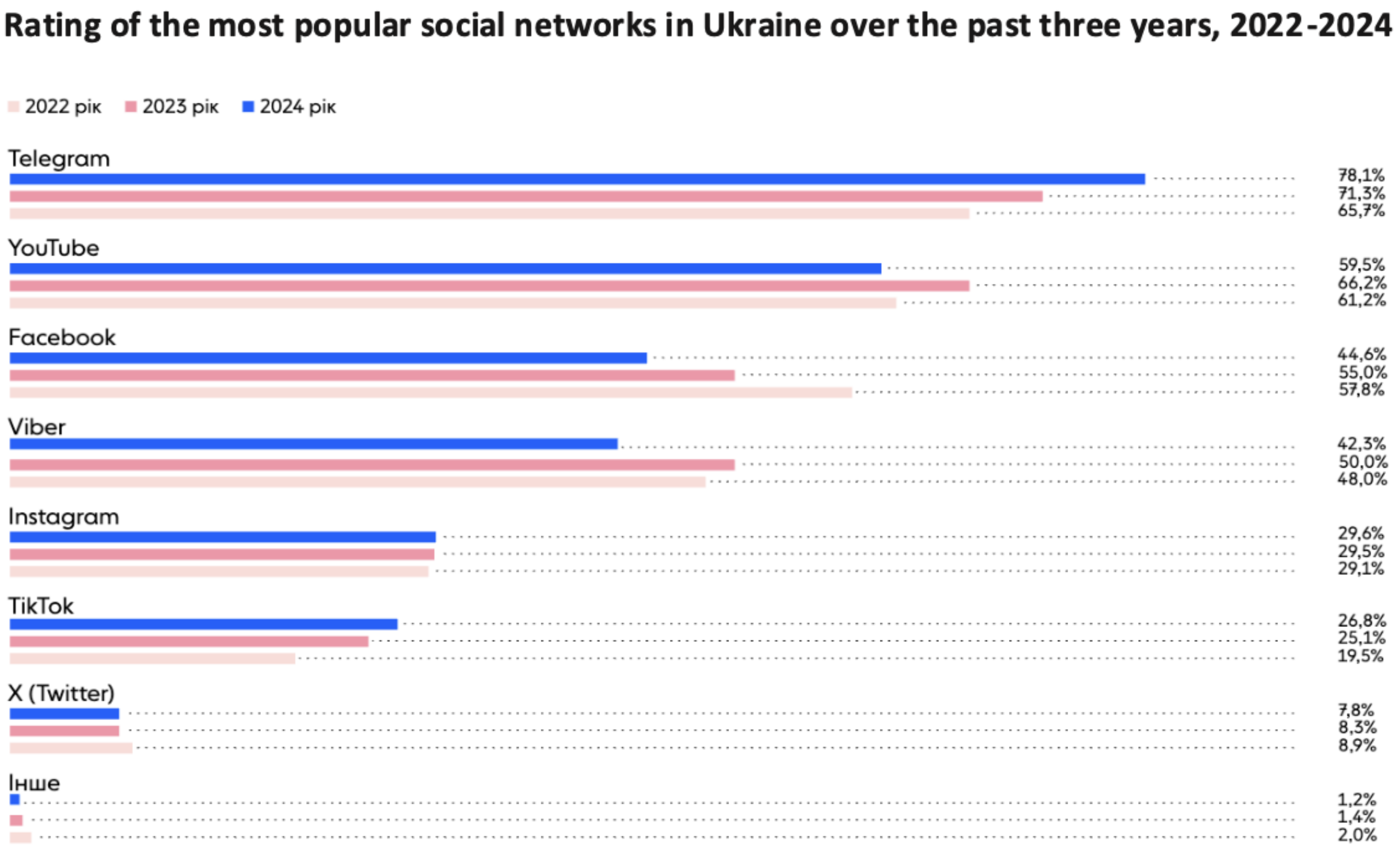
- Models available on Huggingface do not really work
- Good sentiment models allow business to monitor their products and for organisations to monitor social media for public opinion on different topics **at scale**
- Sentiment analysis was shown to predict election outcomes **better than polls**
- Difficulty: annotated data availability and linguistic code-switching complexity of the data

Annual distribution of the articles amount about sentiment analysis in Web of Science, 2013-2024



Quality & Usability Lab

Dataset: why Telegram focus

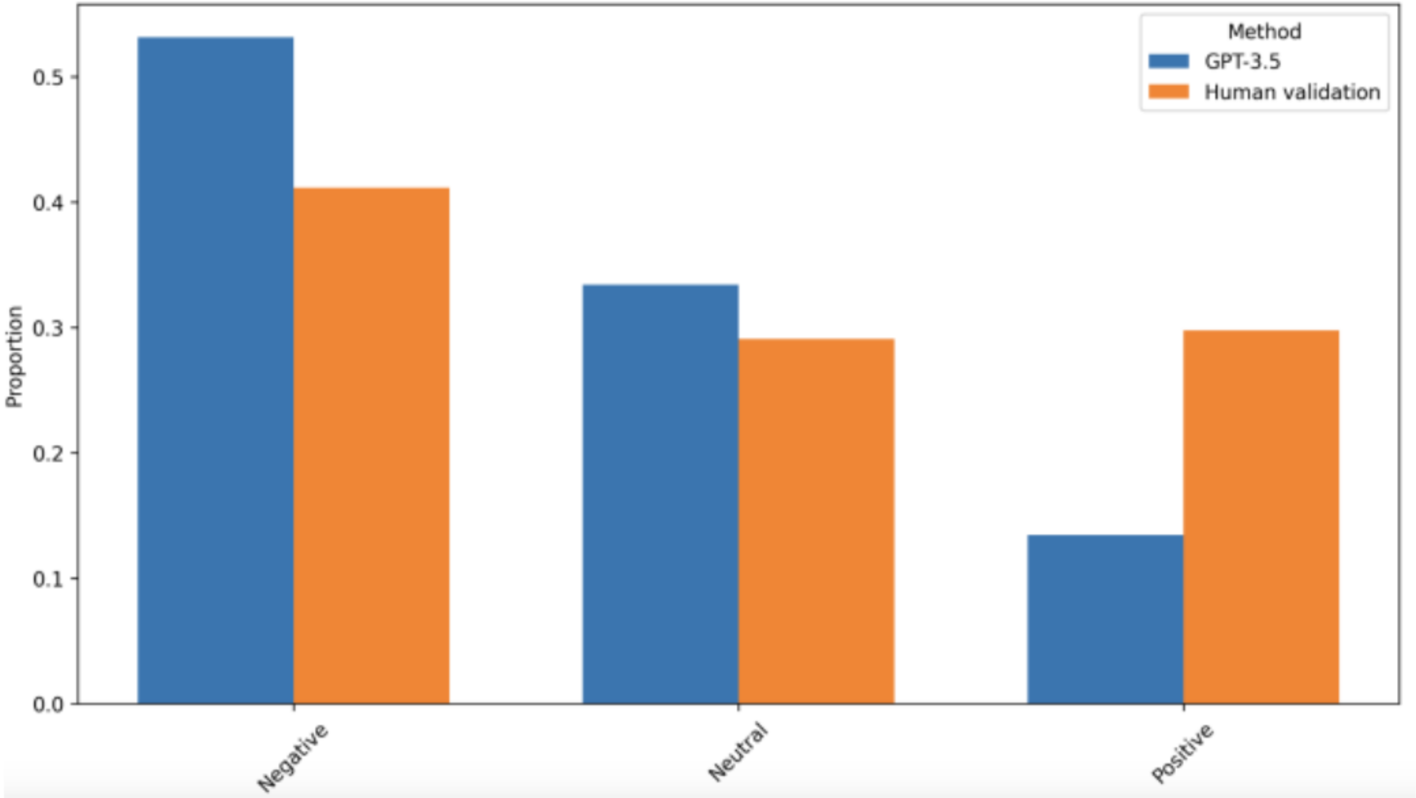


Source:

OPORA social questionary.
Media consumption of Ukrainians: the third year of full-scale war

Dataset: why annotation still matters

The classes distribution of GPT-3.5 versus human validation during the annotation of social media texts about the war in Ukraine



Source:

Ustyianovych, T., Barbosa, D.:

Instant messaging platforms news multi-task classification for stance, sentiment, and discrimination detection.

Dataset: existing datasets

Overview of datasets with Ukrainian social media content and annotated for sentiment analysis

Languages	Volume	Mean length	Classes structure	Sentiment description	Annotation guideline
Russian: 100%	13,114	736	Positive: 52% Negative: 48%	News orientation towards Ukraine	Unknown
Ukrainian: 75% Russian: 25% English: <1%	3,000	143	Positive: 74% Negative: 26%	News orientation towards Ukraine	Unknown
Ukrainian: 62% Russian: 38% English: <1%	7,513	203	Unlabeled: 91% Negative: 3% Neutral: 3% Very Negative: 3%	Text emotion	Manual
Ukrainian: 80% Russian: 20%	564	501	Negative: 57% Positive: 30% Neutral: 13%	Text emotion	Manual
Russian: 98% Ukrainian: 2%	276,309	369	Negative: 53% Neutral: 33% Positive: 14%	Text emotion	GPT-3.5

Not applicable sentiment orientation

Most data is unlabeled

Too small

Mostly in Russian

Quality & Usability Lab

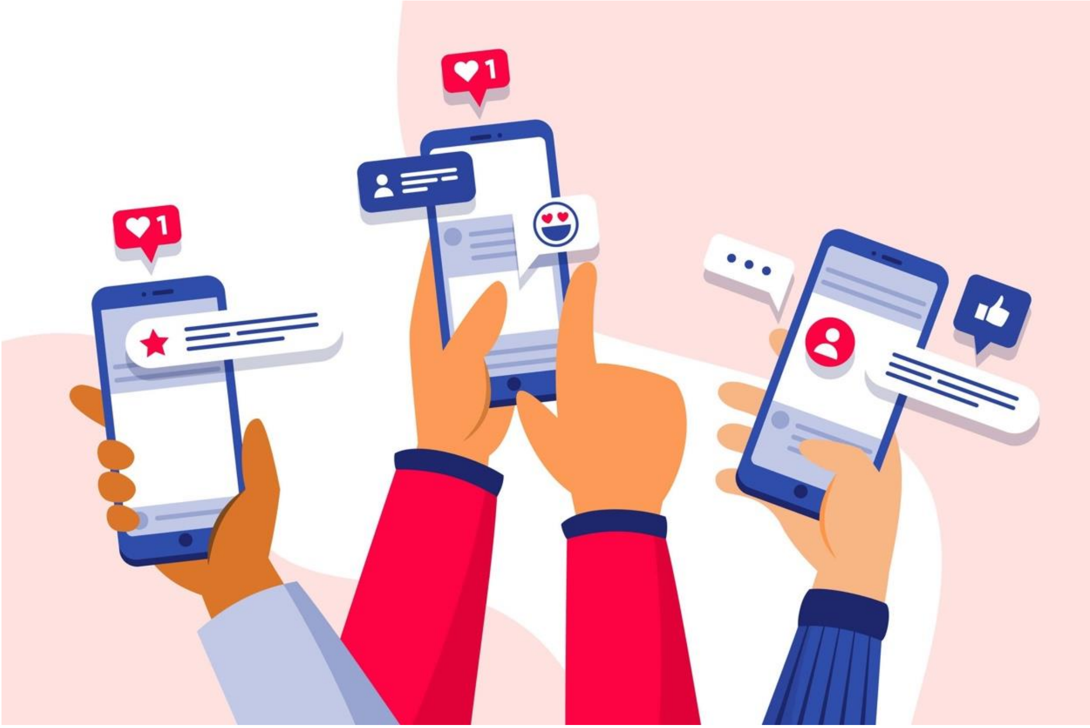
Dataset: COSMOS

- Content dated between February 2022 and September 2024
- The total size of the collected texts is **12,224**

• **7,224 texts** - the volume of the scraped documents (Telegram big news channels and comments)

• We integrated two publicly available datasets: TG samples from D. Baida [1] with 3,000 samples and 1,000 Yakaboo book reviews [2]

• 1,000 product reviews from Hotline.ua



Source:
[1] https://huggingface.co/datasets/dmytrobaida/autotrain-data-ukrainian-telegram-sentiment-analysis
[2] https://github.com/osyvokon/awesome-ukrainian-nlp

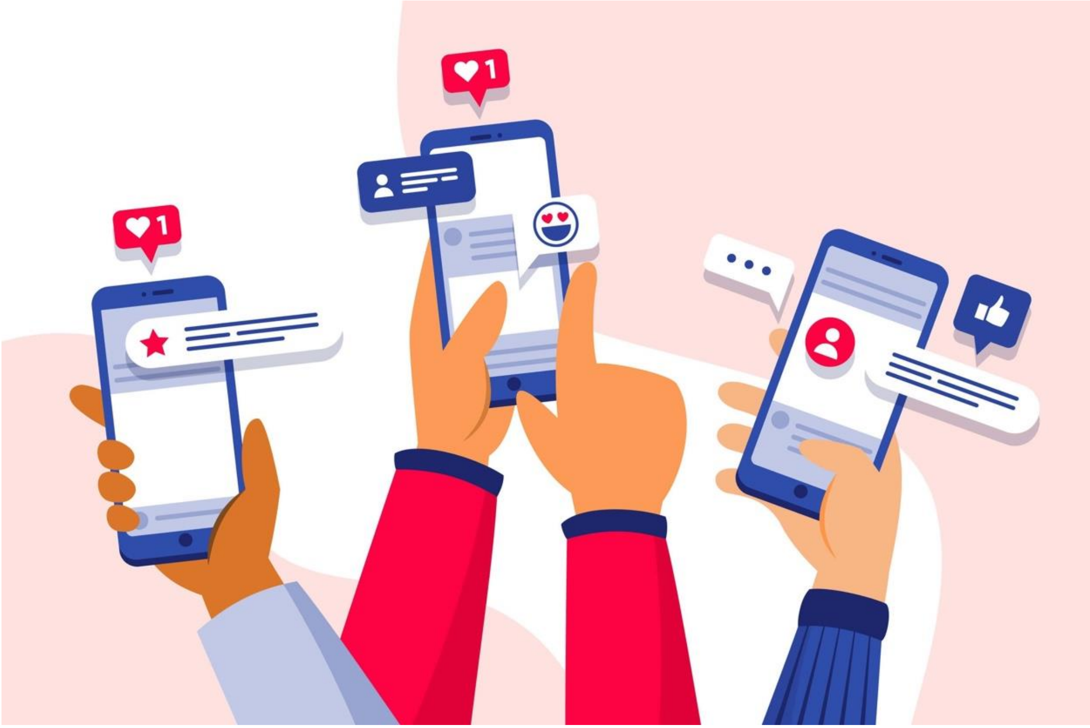
COde-Switched MUltilingual Sentiment for Ukrainian Social media

- Content dated between February 2022 and September 2024
- The total size of the collected texts is **12,224**

• **7,224 texts** - the volume of the scraped documents (Telegram big news channels and comments)

• We integrated two publicly available datasets: TG samples from D. Baida [1] with 3,000 samples and 1,000 Yakaboo book reviews [2]

• 1,000 product reviews from Hotline.ua



Source:

[1] https://huggingface.co/datasets/dmytrobaida/autotrain-data-ukrainian-telegram-sentiment-analysis

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

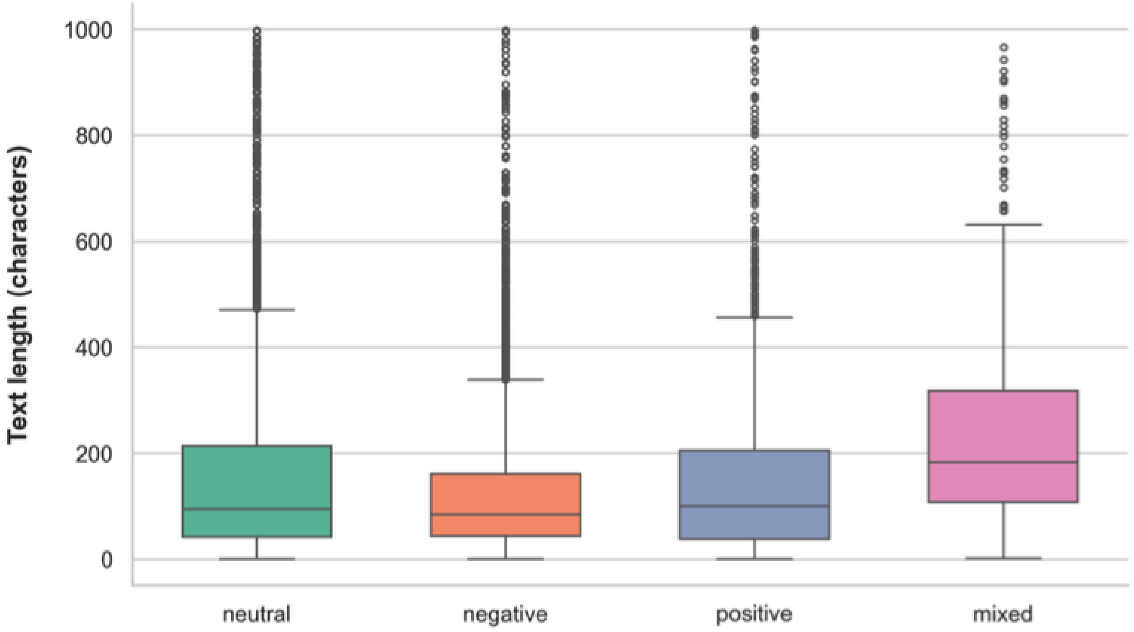
Quality & Usability Lab

Dataset: COSMOS

- Final dataset includes 4 classes:
  - Positive
  - Negative
  - Neutral
  - Mixed

- The resulting dataset includes the following languages in the proportion:
  - 66% - Ukrainian
  - 28% - Russian
  - 6% - code-switched content

Distribution of text lengths (in characters) across sentiment categories in the final dataset



Quality & Usability Lab

Dataset: COSMOS annotations

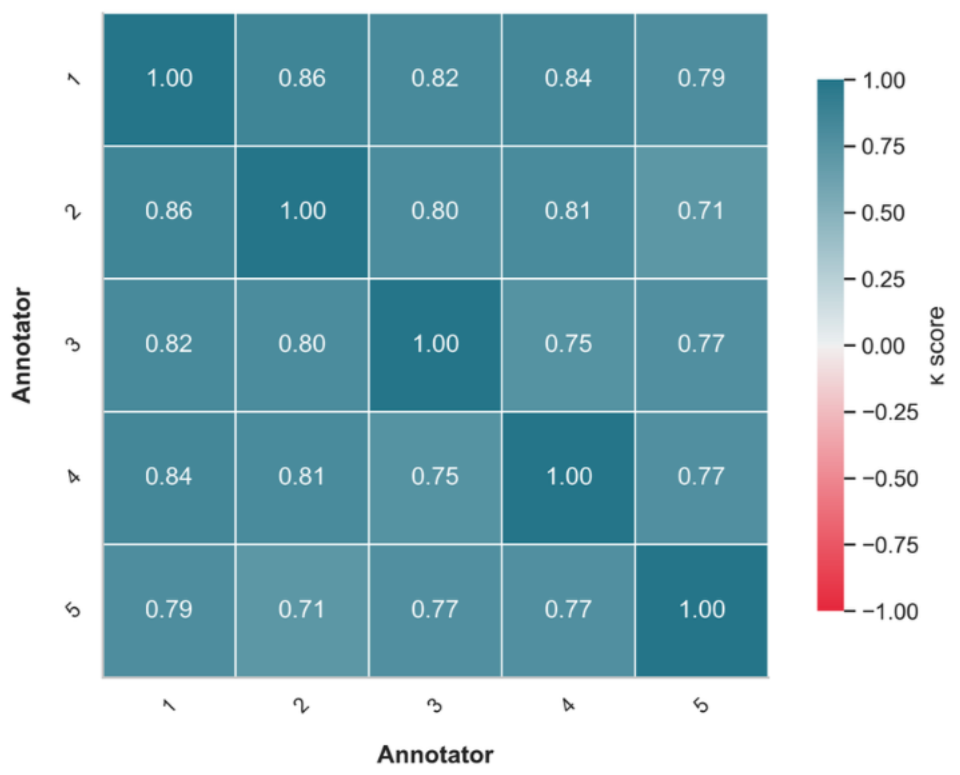
4 The annotators have agreed to join our project (not professional linguists) + 2 first authors (computer scientist and trained linguist)

The average pairwise Cohen's Kappa agreement was $\kappa = 0.79$, indicating substantial reliability

Sentiment distribution of the dataset

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

Inter-annotator agreement matrix for the COSMUS dataset



Quality & Usability Lab

Methods

LLMs and SLMs

SLM calibration analysis

Data augmentation

XAI analysis with LIME scores

Methods

Approaches we tested

- **Prompting baseline**: **GPT-4o*, DeepSeek V2-chat**
- **BERT-like**: **UkrRoBERTa, Modern BERT**

Methods: LLM Baseline

- Across all configurations, prompts in English consistently outperformed Ukrainian prompts for both models
- The performance gap between Ukrainian and English prompting was more pronounced in the zero-shot setting than in the few-shot setting
- GPT-4o consistently outperformed DeepSeek V2-chat across all prompting strategies. This performance difference likely reflects GPT-4o's stronger multilingual capabilities

MacroF1-scores of LLM-based sentiment classifiers across different prompting strategies

Model	Zero-Shot (Ukr)	Zero-Shot (Eng)	Few-Shot (Ukr)	Few-Shot (Eng)
GPT-4o	0.55	0.58	0.61	**0.63**
DeepSeek V2-chat	0.51	0.56	0.58	**0.59**

Methods: Augmentation

- UkrRoberta with word substitution augmentation emerged as the strongest
classifier overall, achieving a macro F1-score of 0.64 on the test set
- The results highlight the importance of selecting an appropriate augmentation strategies based on model architecture and training paradigm

Macro F1-scores of sentiment classification models across different data augmentation strategies

Model	Original	Back-translation	Word substitution
GPT-4o	**0.56**	0.53	0.54
DeepSeek V2-chat	**0.52**	0.50	0.52
mBERT	0.53	0.49	**0.58**
UkrRoberta	0.55	0.52	**0.64**

Quality & Usability Lab

Results

Language Subset	UkrRoberta ECE	mBERT ECE
All Texts	0.17	0.32
Ukrainian-only	0.16	0.40
Code-mixed	0.13	0.35
Russian-only	0.18	0.17

Performance comparison between UkrRoberta and mBERT sentiment classification models

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

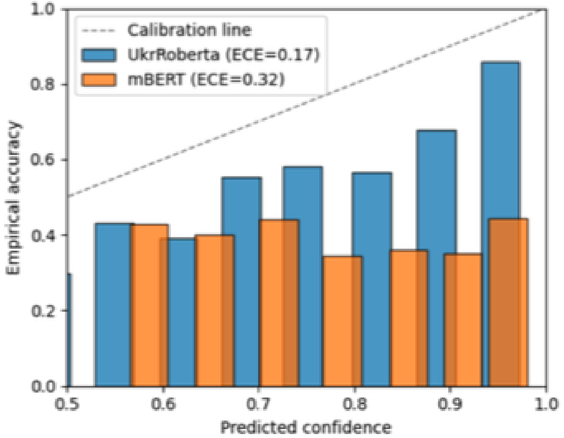
Expected Calibration Error (ECE)

- The key idea here is to evaluate how well a model's predicted probabilities reflect the true likelihood of an outcome — in other words, how calibrated the model is

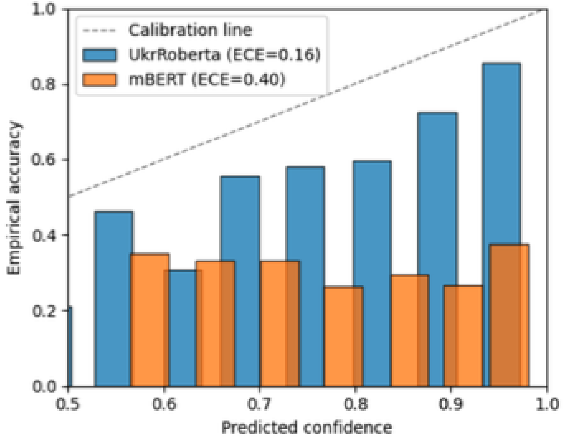
$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|$$

Results: Calibration

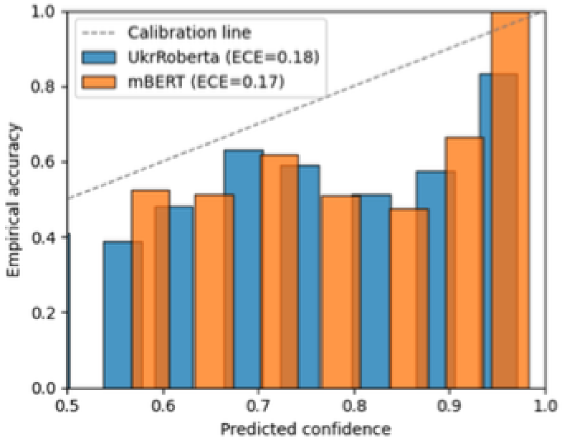
Reliability diagrams for UkrRoberta and mBERT calibration across language subsets



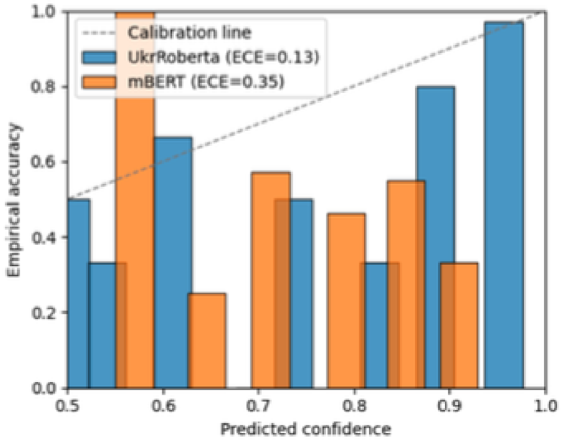
(A) Models overall calibration



(B) Calibration on Ukrainian-only texts



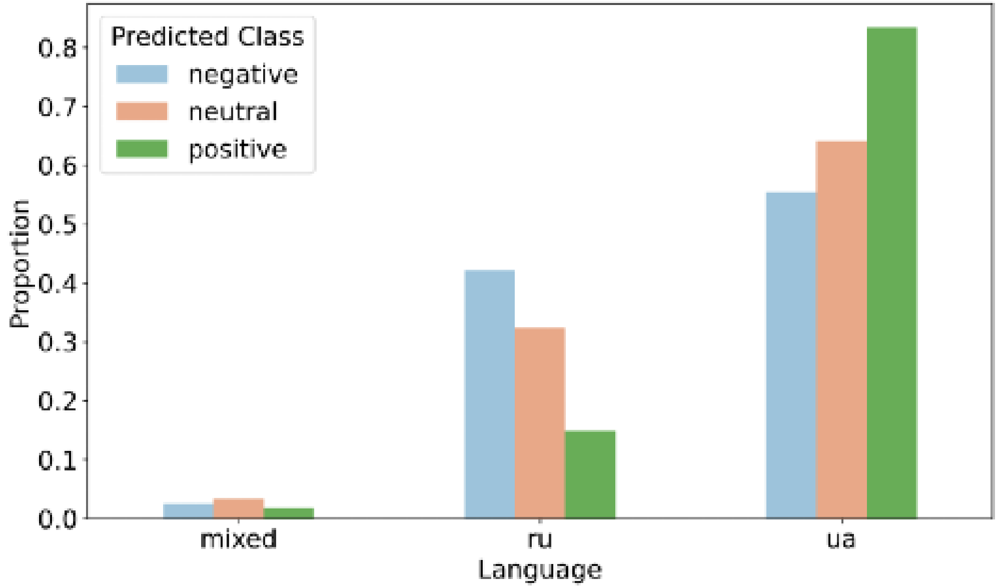
(C) Calibration on Russian-only texts



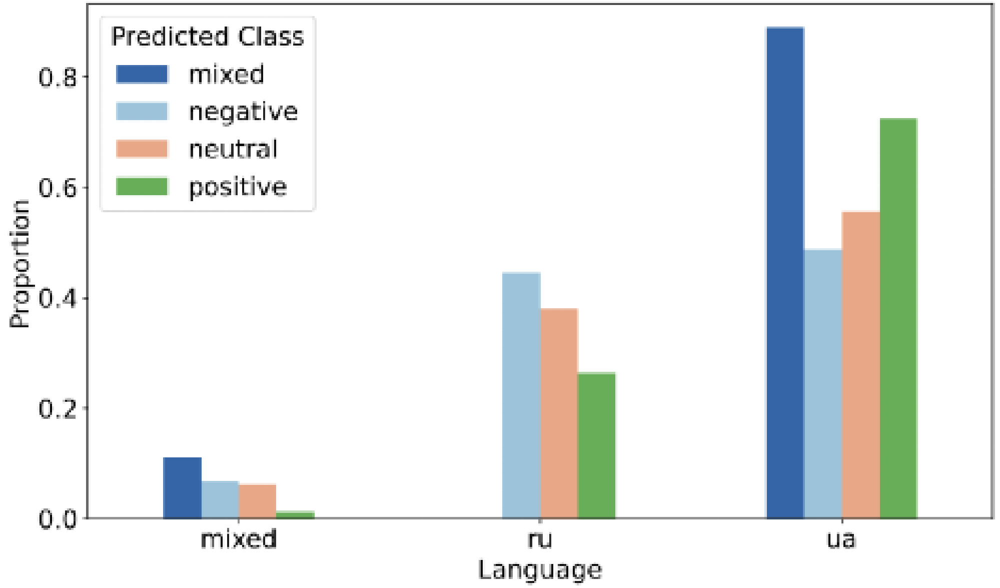
(D) Calibration on code-switched texts

Evaluation: XAI, detecting language bias

Language contribution of the test set to predicted sentiment classes with LIME score



(a) 3-class model LIME analysis



(b) 4-class model LIME analysis

Quality & Usability Lab

Evaluation: XAI

Positive

Ambiguously positive terms: *все*, *воїни*, *ґіґачади*, *вірю*

Irony, sarcasm, and colloquial usage make interpretation harder. Some misclassifications are caused by such cases

Negative

War-related & profane terms: *розбомбленная*, *х уячит*, *обстреливают*, *в ата*, *русня*, *жахливий*

Consistent with emotional intensity. Laughter tokens sometimes misleadingly signal irony or sarcasm.

Neutral

Emotionally neutral terms: conjunctions, generic verbs

Often predicted due to **the absence** of strong sentiment cues, not the presence of neutral ones.

Mixed

Few clear markers; examples include *нах* (strongly negative) and *круто* (positive)

Indicates **weak concept learning** for “mixed” sentiment by the model.

Quality & Usability Lab

Conclusions

1. We developed COSMUS, a publicly available corpus of 12,224 texts covering Ukrainian, Russian, and code-switched content
2. Our experiments demonstrated that targeted word substitution can substantially improve fine-tuning results, while back-translation often degraded model performance
3. Fine-tuned UkrRoberta, combined with word substitution augmentation achieved the best results
4. LIME confirms that **UkrRoberta learns some sentiment-bearing patterns**, but:
•**Fails to fully capture irony and sarcasm**
•**Over-relies on lexical cues** like profanity or named entities
5. Gpt-4o work better than deepseek for Ukrainian social media sentiment annotation, but both cannot reach task-specific model performance.

Links to the dataset and fine-tuned best model

Thank you!

Q&A

veronika.solopova@tu-berlin.de

Yurii Shynkarov, Veronika Solopova, Vera Schmitt