Tuning of language models in Eastern European languages on Twitter/X

Tomáš Filip[1,†], Martin Pavlíček[1,†] and Petr Sosík[1,2,*,†]

[1]*Institute for Research Applications of Fuzzy Modeling, University of Ostrava, 30. dubna 22, Ostrava, 70200, Czech Republic*
[2]*Institute of Computer Science, Faculty of Philosophy and Science, Silesian University in Opava, Bezručovo náměstí 1150/13, Opava, 746 01, Czech Republic*

Abstract

We address the problem of fine-tuning large language models (LLMs) for sentiment analysis on Twitter/X in underrepresented Eastern European languages (Czech, Slovak, Polish, and Hungarian). We study the influence of a number of experimental settings on the efficiency of fine-tuning in two groups of LLMs: transfer-learning models (BERT, BERTweet or XLM-T, the latter two pre-trained on a Twitter corpus) and popular mid-sized universal models (Llama, Mistral). We show that adapter fine-tuning with as few as ≈ 600 tweets improved scores of our universal models to the level previously reported by Twitter/X-specialised models on popular datasets, while our transfer-learning models performed worse. We also show that, despite previous successful experiments with multilingual models, translating from underrepresented languages into English still improves the results of all models tested. Several other factors that influence the success of fine-tuning are also included in the study.

Keywords

Large language model, Sentiment analysis, Twitter, Eastern-European language, Russo-Ukraine conflict, Llama, Mistral, BERTweet, BERT, XLM-T, GPT-4

1. Introduction

Sentiment analysis is one of the most common topics in natural language processing, with rapidly emerging techniques [1]. Recently, machine learning methods, especially large language models (LLM), have been considered the state of the art on sufficiently large training datasets. As end-user deployment of language models is now common and affordable, their performance in underrepresented languages is becoming important.

This paper focusses on fine-tuning LLMs for sentiment analysis in Eastern European languages (Czech, Slovak, Polish, and Hungarian) belonging to the so-called Visegrád (V4) group. As a case study, we chose the topic of the Ukraine war crisis on Twitter/X, providing a large textual corpus with rich sentiment polarity. This topic is also the target of intensive cyberbullying attacks and, simultaneously, a crucial source of Open Source Intelligence (OSINT), further underlining its relevance. The novelty aspects:

- Twitter/X studies in Eastern European (EE) languages are rare in LLM-based sentiment analysis, and we are not aware of any studies focussing on the Russo-Ukraine conflict.
- The aspects of the tunability of various LLMs on Twitter/X (or similar) EE data have not been adequately researched.
- The performance of mid-sized or large models (Llama, Mistral, or GPT-4) versus transfer learning models (BERT, BERTweet, RoBERTa) in Twitter/X-based tasks has been poorly studied, with very few exceptions, such as [2, 3].

25th Conference on Information Technologies – Applications and Theory (ITAT)
*Corresponding author.
†These authors contributed equally.
tomas.filip@osu.cz (T. Filip); martin.pavlicek@osu.cz (M. Pavlíček); petr.sosik@osu.cz (P. Sosík)
0009-0001-4386-0620 (T. Filip); 0000-0003-1429-2668 (M. Pavlíček); 0000-0001-7624-3816 (P. Sosík)

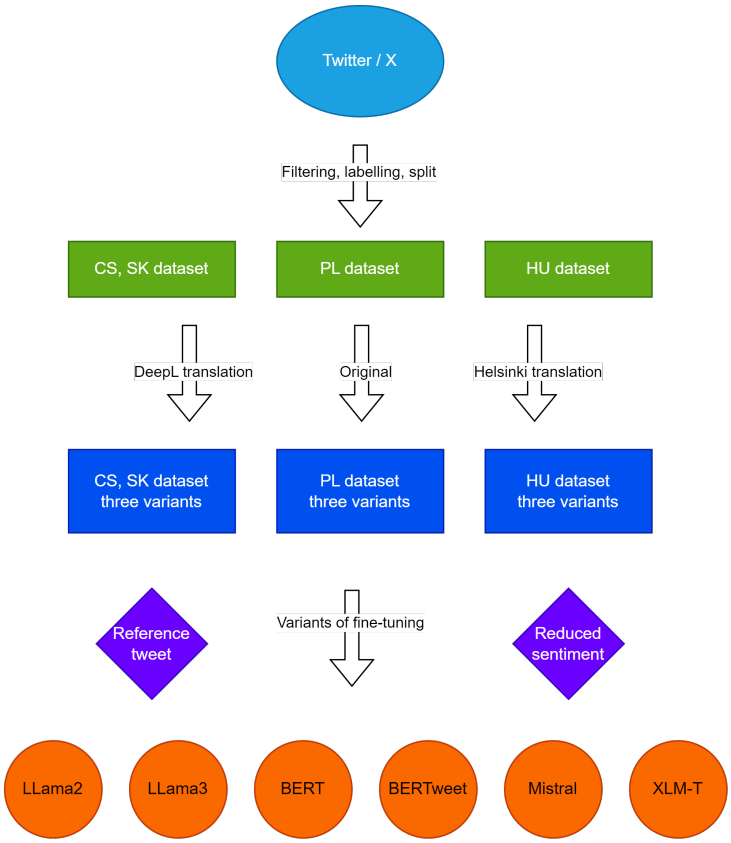

Figure 1: Experimental pipeline overview. The downloaded dataset was split into three language-specific parts. Three version versions of translation (Helsinki, DeepL, none) were prepared, obtaining 9 individual datasets. The tested models were fine-tuned in four variants, combining classification into two/three classes and training with/without reference tweets.

We downloaded and annotated three monolingual datasets (CS/SK, PL, HU) from Twitter/X. The dataset was used to fine-tune three transfer learning models (BERT, BERTweet, XLM-T) and three mid-sized LLMs (Llama 2, Llama 3, Mistral) in a number of experimental settings illustrated in Fig. 1. The training objective was the sentiment polarity towards either Ukraine or Russia. We evaluated the influence of various settings, such as the size of the dataset, the translation into English, or the presence of the reference tweet (the one to which the tweet reacted) on the efficiency of fine-tuning. The key findings are as follows.

- Fine-tuning with as few as $\approx$ 600 tweets in underrepresented Eastern European languages improved the F1 score of the Llama and Mistral models by 30–40%, reaching the level of specialised models on Twitter/X benchmarks.
- Fine-tuned general mid-sized LLM such as Llama or Mistral significantly outperformed equally fine-tuned transfer learning models (BERTweet, XLM-T) pre-trained on a large Twitter/X corpus.
- All models (including multilingual XLM-T or GPT-4) performed best when fine-tuned on a dataset translated into English by DeepL.
- Unsurprisingly, in-context learning did not help the small- and mid-sized models, but neither the context of the reference tweets improved the fine-tuning.

The rest of the paper is organised as follows. Section 2 briefly resumes sentiment analysis in texts, with a focus on Twitter/X datasets. Section 3 describes the construction of our dataset, followed by Sec. 4 that outlines the experimental settings. Section 5 contains an overview of the results, which are then discussed in more detail in Sec. 6. Section 7 provides an ablation study that focusses on the impact of selected experimental variables. Finally, Section 8 summarizes the results.

2. Background

With the rapid growth of social networks and e-commerce, sentiment analysis has emerged as one of the fastest-growing research areas in computer science. To capture sentiment with greater granularity, Hu and Liu [4] introduced the concept of aspect-based sentiment analysis (ABSA) in their foundational work, which has since inspired numerous follow-up studies. A comprehensive review of recent developments in NLP-based sentiment analysis is provided by Jim et al. [1].

Recent progress in ABSA has been significantly driven by the integration of large language models (LLMs).[1] For example, Zhang et al. [5] proposed a generative framework that formulates ABSA as a text generation problem, offering a flexible alternative to traditional classification approaches. Building on the strengths of instruction-based learning in LLMs, Scaria et al. [6] introduced the InstructABSA model, which leverages task instructions to improve performance. Periodic survey studies, such as that by Brauwers and Frasincar [7], continue to provide structured overviews of the evolving ABSA landscape.

Sentiment classification can be challenging in Twitter/X data due to the lack of explicit context and specific style. TweetEval benchmark [8] evaluated models that analyse sentiment in tweets on detection tasks of emotion, irony, hate speech, offensive language, stance, emoji prediction and sentiment analysis. The TweetEval leaderboard on GitHub lists BERTweet [9] as the current SoTA model, closely followed by TimeLM-21. The family of TimeLM models [10] reflects the current context problem by periodic updates with tweet datasets, and outperformed BERTweet in many tasks.

Barbieri et al. [11] expanded the focus on multilingual tweet analysis and presented a unified tweet benchmark in eight languages (UMSAB). The paper also introduced the XLM-Twitter model (XLM-T) developed by pre-training the XLM-R [12] using 198M multilingual tweets. XLM-T was further fine-tuned in UMSAB, and the resulting model was named XLM-T Sentiment. Barreto et al. [13] studied, among other topics, the performance of BERT, RoBERTa and BERTweet in Twitter ABSC tasks.

Krugmann et al. [2] compared the performance of established transfer learning models (BERT, BERTweet, RoBERTa) with recent LLM (GPT-3.5, GPT-4, and Llama 2) on Twitter/X data, with the superiority of the latter. In contrast to these results, Stigall et al. [3] presented a fine-tuned model $\text{EmoBERT}_{\text{Tiny}}$ for emotion and sentiment classification tasks and reported its superiority over non-tuned Llama-2-7B-chat and Mistral-7B-Instruct across all metrics. These and other authors also reported on the domain sensitivity of the models.

Finally, of many existing sentiment studies on Russia–Ukraine war on social networks, we mention two. An evaluation of traditional ML models (logistic regression, decision trees, random forests, SVMs etc.) on Twitter data was provided in [14]. A deep learning approach combining multi-feature CNN with BiLSTM was applied in [15] to an analogous task. Both studies relied on monolingual English datasets.

3. Dataset construction

Our data were collected using the academic Twitter/X API during the period 4/2/2023 to 20/5/2023. Filtering by languages (Czech/Slovak, Polish, Hungarian), and keywords (Ukraine, Russia, Zelensky, Putin) resulted in 34,124 relevant tweets split into three monolingual parts according to the language. There was no filter available for Slovak so it was mixed with Czech. In every monolingual dataset, we manually annotated a random subset of tweets by their sentiment toward Ukraine or Russia, keeping the classes roughly balanced. Certain class imbalance resulted from the lack of relevant tweets neutral to a given aspect. To avoid annotation bias, the annotators followed the principles of the CAMEO[2] conflicting topic codebook, and the annotated tweets were cross-validated among the annotators. The annotated datasets are not the same size (see Table 1), to study the impact of the size on the models'

[1] https://paperswithcode.com/task/aspect-based-sentiment-analysis
[2] http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

Table 1
Size of language-specific subdatasets (No. of tweets)

Lang.	Aspect	Total	Pos.	Neutral	Neg.
cs/sk	Ukraine	1638	632	447	559
cs/sk	Russia	1716	579	537	600
pl	Ukraine	640	205	263	172
pl	Russia	570	202	164	204
hu	Ukraine	628	202	203	223
hu	Russia	556	181	145	230

Table 2
Language models used in the experiments.

Model	Total params	Tuned params	Paper	Web page
BERT base	110M	110M	[18]	https://huggingface.co/google-bert/bert-base-uncased
BERTweet large	355M	355M	[9]	https://huggingface.co/vinai/bertweet-large
XLM-T Sentiment	279M	279M	[11]	https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
Llama-2 7B	6.7B	4M	[19]	https://huggingface.co/meta-llama/Llama-2-7b-hf
Llama-3 8B	8B	3.4M	[20]	https://huggingface.co/meta-llama/Meta-Llama-3-8B
Mistral 7B	7.2B	3.4M	[21]	https://huggingface.co/mistralai/Mistral-7B-v0.1

performance. Each annotated dataset was split into a training set (75 %) and a testing set (25 %). The datasets are available in the supplementary data on GitHub; see the link in Conclusions.

4. Methods

Language models

The models we tested (Table 2) belong to two categories: (i) transfer learning models popular in the ABSA literature and in the TweetEVAL and UMSAB benchmarks: BERT, BERTweet, and XLM-T. The latter two have been pre-trained on large Twitter/X corpuses. As we intended to study the tunability of universal models, we did not use language-specific variants as the PolBERT[3], huBERT[4] or the SlovakBERT[5]. (ii) Mid-sized open-source models (up to 10B parameters) which are fine-tunable on limited end-user GPU hardware: Llama-2 7B, Llama-3 8B, and Mistral 7B. Recent studies such as [2, 3, 16] point out missing studies on ABSA using these and similar models. Furthermore, ChatGPT-4 [17] was used as a reference model for tweet classification.

Translation

When applying pre-trained LLMs to datasets in underrepresented languages, some sources such as [22, 23] report better results with machine translation to English, while others rely on follow-up training or fine-tuning in original languages [11, 12]. To compare the effectiveness of both approaches, the annotated datasets were used for both training and testing in three different language modes:

- translated to English using the Helsinki Neural Machine Translation System[6];

[3]https://github.com/kldarek/polbert
[4]https://huggingface.co/SZTAKI-HLT/hubert-base-cc
[5]https://huggingface.co/gerulata/slovakbert
[6]https://huggingface.co/Helsinki-NLP

- translated to English using the DeepL API[7];
- no translation, original languages (CS/SK, PL, HU).

Training

We trained each decoder model by using a tweet as input and generated a single output token. The loss function was the cross-entropy between the generated token and the ground truth label. Each model in Table 2 in combination with each translation mode was fine-tuned on each language-specific training set (not their combination). For Llama 2/3 and Mistral we used the PEFT adapter-based technique [24] using the Python PEFT library[8]. The number of tuned parameters varied between 3.5–4 million. The training was run for 10 epochs on all models. The learning rate was set to $3e^{-4}$, batch size was 4. The learning rate schedule was linearly growing to maximum during warm-up (the first 100 iterations) and then linearly decreasing towards zero. The remaining hyperparameters were library-default. All metrics were calculated at the best checkpoint of the model. Both training and inference were run on a server 2 x 2060 RTX (8GB) for smaller BERT-derived models, and another server with 2 x NVIDIA V100 (32GB) for larger models.

Inference

After fine-tuning in a specific language, all models in 2 were prompted the same way using the testing set in the same language. The experiments were carried out with and without the use of the reference tweet (to which the classified tweet reacted). We used a simple English prompt in all experiments:

tweet: {tweet}
The sentiment of the tweet towards {aspect} is...

For GPT-4 we did not use fine-tuning but instead applied in-context instruction learning (ICL), that is, expanding the prompt with context information related to the question asked. The expanded prompt can be found in the online Appendices to the paper; please follow the link in the "Supplementary material" section.

5. Results

We conducted an extensive series of tweet sentiment classification experiments that varied in the following settings:

- sentiment aspect (Russia/Ukraine)
- language of the tweet (CS/SK, HU, PL)
- language model (BERT, BERTweet, XLM-T, Llama 2, Llama 3, Mistral, GPT-4)
- tweet translation (DeepL, Helsinki translator, none)
- positive/neutral/negative classification, or only positive/negative
- the presence of a reference tweet

Standard metrics were used to evaluate the results: accuracy and macro-averaged recall, precision, and F1 score [25]. The macro-averaged F1 was chosen as our primary evaluation measure due to its balanced assessment for the evaluation of model performance across multiple classes (negative, neutral, positive). Unless stated otherwise, tables and graphs show results for *positive/neutral/negative* sentiment classification. With the exception of ChatGPT-4, all results were obtained without using reference tweets. The complete results are contained in the supplementary data on GitHub; see the link in Conclusions.

[7]https://www.deepl.com/translator
[8]https://huggingface.co/docs/peft

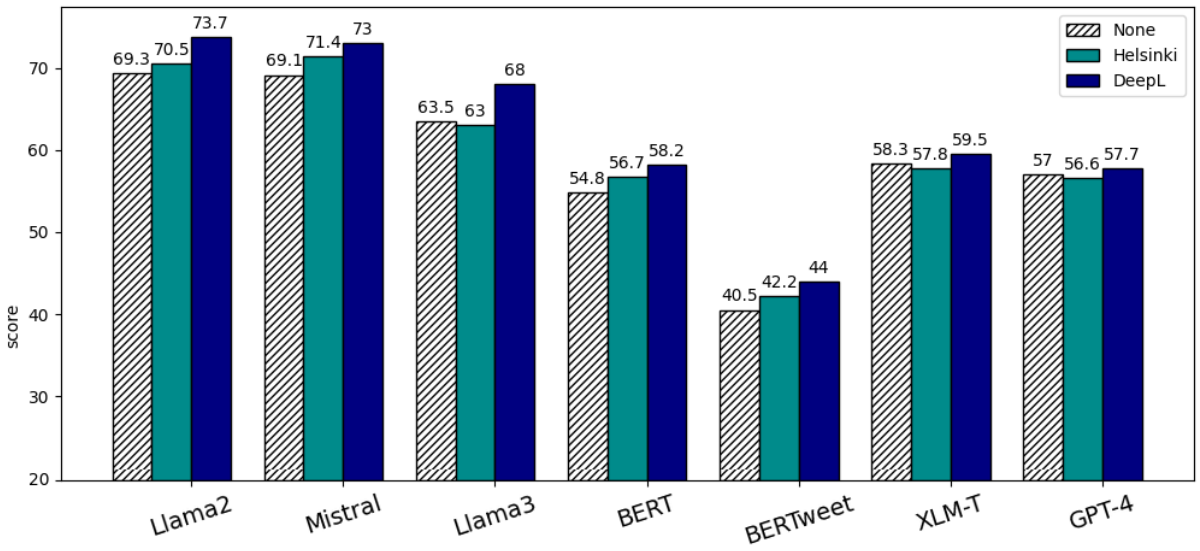

Figure 2: Macro-averaged F1 score by language models and translation.

Results by translation

Figure 2 summarises the main results organised by language models and type of translation. Concerning the performance of individual models, surprisingly, Llama 3 scored approx. 6% F1 worse than Llama 2 and BERTweet large performed worse than BERT base; perhaps pre-training on older tweets could have affected tunability of BERTweet to a newer context. Neither did XLM-T reach the level of the larger models, although it was pre-trained on a large multilingual tweet corpus. The order of magnitude larger model size seems to be the prevailing factor. Finally, all models benefitted from the DeepL translation. Therefore, the remaining results included in the paper are restricted to DeepL-translated datasets.

Results by languages

Figure 3 summarises experiments on individual languages using DeepL translation. Quite surprisingly, almost all models performed poorer for the Polish language. These results do not correlate with the support of the individual languages datasets (see Table 1) nor with the type of translation, and cannot be attributed to pre-training either (e.g., GPT-4 performed well in Polish on the MMLU benchmark [17]). The results were similar also for the vanilla models tested; see Table 3. A detailed analysis showed that many positive Polish tweets were classified as negative by the models. These tweets contained more complex thoughts about the historical interconnection of Poland with Ukraine. Some examples of misclassified tweets can be found in appendices in the supplementary material, please see the link in Conclusions.

6. Discussion

Relation to the SoTA

Our focus on underrepresented EE languages does not allow direct comparison with popular Twitter/X benchmarks, and the following figures provide only an approximate picture. The TweetEVAL leaderboard [8] marks TIMELM-21 as the SoTA model with macro-averaged recall 73.7 for three-valued ABSA, followed by BERTweet with recall 73.4. Our best macro-averaged result (Llama 2, translation by DeepL, averaged over all aspects and languages) provided the F1 score 73.7. Our task is on the one hand much narrower than TweetEVAL. On the other hand, TweetEVAL is monolingual and BERTweet was trained on 850M English tweets, while we fine-tuned our models using three datasets with a few hundreds of tweets in underrepresented languages.

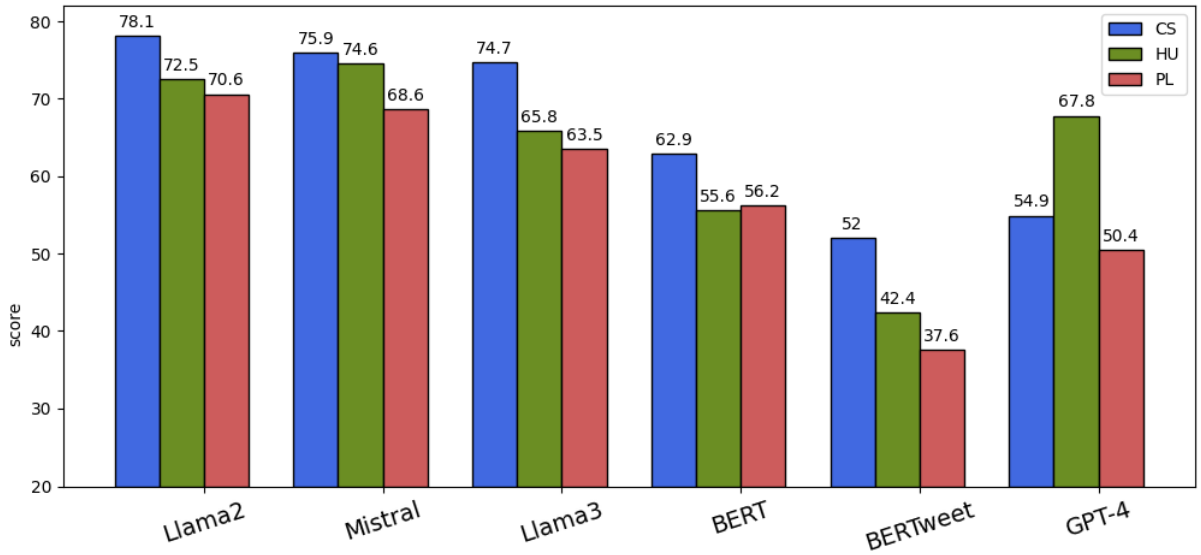

Figure 3: F1 score by models and languages of tweets for *positive/neutral/negative* classification, macro-averaged over both aspects UA/RU.

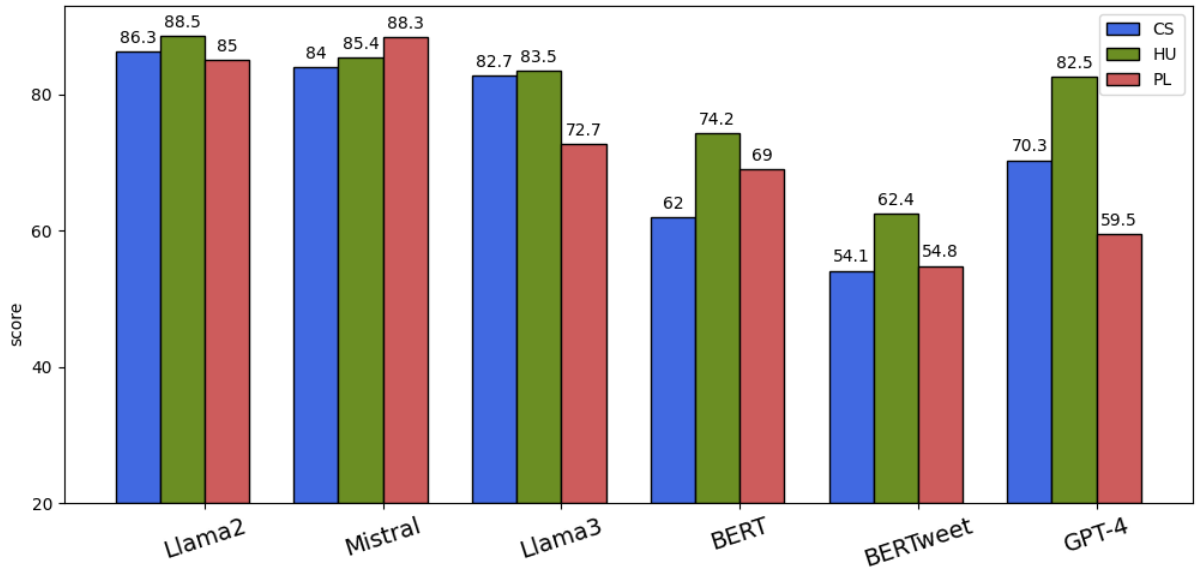

Figure 4: F1 score by models and languages of tweets for *positive/negative* classification, macro-averaged over both aspects UA/RU.

The UMSAB Twitter benchmark [11] reports XLM-Tw Multi as the best model with an F1 score of 69.4, macro-averaged in eight languages. Again, this task is wider than ours, but XLM-Tw Multi used a much larger fine-tuning dataset; therefore, we cannot provide an exact comparison.

Size of the training sets

The support of the CZ/SK training set was approximately three times that of HU or PL which were almost equal. This imbalance allowed for some interesting observations. In the simpler task of two-valued classification, almost all fine-tuned models returned scores irrelevant to the language, implying that the training sets with about 600 tweets were sufficient to bridge the language differences. However, in the case of three-valued classification, the CZ / SK dataset was favoured by all fine-tuned models. Hence, for this harder task, the smaller HU/PL training set was insufficient. The effect was stronger for smaller models (BERT, BERTweet), confirming the multiplicative joint scaling law for LLM fine-tuning

[26].

Model and human bias

In the context of the current situation where Russia is described as the aggressor, human annotators who know more about the context may tend to see the situation in terms of cause and effect, and therefore their sentiment determination is usually biased differently than the models [27]. In particular, LLMs struggled with tweets neutral (or positive) to a given aspect but generally negative, for example, addressing bombing, war, attack. Models such as Llama 2 or Mistral showed significantly lower precision and recall for the neutral class than for the negative or positive one.

Scores for individual classes

All experiments in Section 5 used macro-averaged recall, precision and F1 scores, since the scores were mostly similar for all classes, with a few exceptions. In particular, in Hungarian, the recall of the positive class was often approximately 10% lower than that of the negative class, and the trend was opposite in precision, meaning that the models tended to classify Hungarian tweets more negatively than human annotators. This might possibly be due to the fact that the overall ratio of negative samples in the Hungarian dataset was a bit higher than in the other languages.

Non effective in-context learning

When employing small, computationally inexpensive models, in-context learning (ICL) often entails notable trade-offs. Due to their more limited representational capacity, these models may be unable to leverage ICL effectively. Another contributing factor may be insufficient pre-training alignment with the target domain or topic. Furthermore, the additional complexity introduced by ICL can increase task ambiguity in aspect-based sentiment analysis (ABSA). Fine-tuning may also override any marginal gains that ICL might provide. To rigorously identify the primary factors underlying the lack of ICL effectiveness, further fine-grained experimental analyses are required.

7. Ablation study

In this section, we discuss the contribution of several components of the experimental pipeline to the classification performance.

Reference tweet use

The reference tweet was always used in the in-context prompt for GPT-4 as it improved its performance (data not shown). For all other models, reference tweets slightly worsened the macro-averaged F1 score (e.g., Bert by 4%, XLM-T by 2%, Llama 2 by 0.5%, Llama 3 by 0.8%, Mistral by 2.5% in the case of *positive/neutral/negative* classification). Therefore, we agree with [28] that, while smaller models rely substantially on semantic priors from pre-training, large models can override them by contradicting exemplars contained in the prompt.

Fine-tuning and in-context learning

To compare these two approaches for Twitter/X task adaptation, we evaluated models Llama 2, Llama 3, Mistral, and GPT-4 in the vanilla version, i.e., without fine-tuning and in-context learning, respectively. The study was restricted to the case of DeepL translation and positive/neutral/negative classification. Table 3 shows that fine-tuning improved the F1 score of Llama 2/3 and Mistral mainly by 20–40% over the vanilla versions, while GPT-4 benefited from the ICL by about 10%.

Table 3
Macro-averaged F1 scores of vanilla (no fine-tuning or in-context learning) and fine-tuned versions of selected models. Setting: DeepL translation, no reference tweets.

Lang.	Target	Llama 2		Llama 3		Mistral		GPT-4	
		Vanilla	Tuned	Vanilla	Tuned	Vanilla	Tuned	Vanilla	ICL
cs	ua	38.5	76.9	37.8	72.2	52.5	72.3	48.8	57.9
cs	ru	40.2	79.2	47.1	77.1	40.4	79.4	49.8	51.8
hu	ua	41.7	70.3	44.7	58.9	50.7	73.4	55.9	66.7
hu	ru	53.9	74.6	47.4	72.6	43.3	75.8	60.5	68.8
pl	ua	33.7	71.1	24.3	62.7	48.2	68.3	39.8	45.3
pl	ru	34.8	70.0	35.7	64.3	35.6	68.9	46.1	55.4

Translation into English

In the overwhelming majority of settings (see Fig. 2 and the supplementary material), all LLMs performed better when fine-tuned and tested on English-translated datasets, and the DeepL translator gave better results than the Helsinki translator. The improvement in the macro-averaged F1 score in all models was 0.8% for the Helsinki translator and 3.1% for the DeepL. DeepL translation improved the F1 score by 1.2% even for the multilingual XLM-T sentiment model. In the supplementary material, we also provide the comparison of the original tweets with both translated versions, to ensure that the classification differences were caused by the quality of the translation and not by a systematic bias of sentiment caused by the translator.

8. Conclusion

We addressed the fine-tuning of large language models for sentiment analysis tasks on Twitter/X in underrepresented Eastern-European languages. We manually annotated a Twitter/X-based dataset related to the Russo-Ukrainian conflict, narrowed to the V4 (Czech Republic, Slovakia, Poland, Hungary) language space. The dataset was used to fine-tune six language models (BERT, BERTweet, XLM-T, Llama 2/3, Mistral) used frequently for sentiment analysis. The tuning was done separately for each language in several variants, using either the original tweets or the English translation with the Helsinki or DeepL translator. Furthermore, GPT-4 (with or without in-context learning) was used as a reference model. The results were evaluated using standard metrics, mostly F1.

We demonstrated that adapter fine-tuning, even with as few as hundreds of samples in underrepresented languages, was able to draw the model's attention to the desired aspects and also to balance language and culture differences (at least for most models). Experiments have shown that, despite previous successful experiments with multilingual models [11, 12], translating from underrepresented languages into English still improves the fine-tuning of all models tested in a wide variety of experimental settings. However, neither the instruction in-context learning nor the enrichment of fine-tuning with the context of reference tweets improved the results. Finally, our experiments also confirmed that the success of fine-tuning depends on the model and the task, as reported by other studies such as [26].

Acknowledgments

This article was produced with the financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition, and under the: Biography of Fake News with a Touch of AI: Dangerous Phenomenon through the Prism of Modern Human Sciences project no.: CZ.02.01.01/00/23_025/0008724 via the Operational Programme Jan Ámos

Komenský. It was also supported by the Silesian University in Opava under the Student Funding Plan, project SGS/9/2024.

Supplementary material and data

https://github.com/zrecorg/zrec-paper-a-study-on-eastern-european-v4-languages

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT4-o), Writefull (Overleaf integration) and DeepL for following - Text Translation, paraphrase and reword, improve writing style, grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, M. Mridha, Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review, Natural Language Processing Journal (2024) 100059.

[2] J. O. Krugmann, J. Hartmann, Sentiment analysis in the age of generative ai, Customer Needs and Solutions 11 (2024) 3.

[3] W. Stigall, M. A. Al Hafiz Khan, D. Attota, F. Nweke, Y. Pei, Large language models performance comparison of emotion and sentiment classification, in: Proceedings of the 2024 ACM Southeast Conference, 2024, pp. 60–68.

[4] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[5] W. Zhang, X. Li, Y. Deng, L. Bing, W. Lam, Towards generative aspect-based sentiment analysis, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 504–510.

[6] K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, C. Baral, InstructABSA: instruction learning for aspect based sentiment analysis, arXiv preprint arXiv:2302.08624 (2023).

[7] G. Brauwers, F. Frasincar, A survey on aspect-based sentiment classification, ACM Computing Surveys 55 (2022) 1–37.

[8] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, L. Neves, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, in: Proceedings of Findings of EMNLP, 2020, pp. 1644––1650.

[9] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: a pre-trained language model for English tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.

[10] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, J. Camacho-Collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2022, pp. 251–260.

[11] F. Barbieri, L. E. Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 258–266.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[13] S. Barreto, R. Moura, J. Carvalho, A. Paes, A. Plastino, Sentiment analysis in tweets: an assessment study from classical to modern word representation models, Data Mining and Knowledge Discovery 37 (2023) 318–380.
[14] G. K. Wadhwani, P. K. Varshney, A. Gupta, S. Kumar, Sentiment analysis and comprehensive evaluation of supervised machine learning models using Twitter data on Russia–Ukraine war, SN Computer Science 4 (2023) 346.
[15] S. Aslan, A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine–Russia conflict, Applied Soft Computing 143 (2023) 110404.
[16] N. Mughal, G. Mujtaba, A. Kumar, S. M. Daudpota, Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis, IEEE Access (2024).
[17] J. Achiam, S. Adler, S. Agarwal, et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[20] AI@Meta, Llama 3 model card, 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[22] M. Araújo, A. Pereira, F. Benevenuto, A comparative study of machine translation for multilingual sentence-level sentiment analysis, Information Sciences 512 (2020) 1078–1102.
[23] V. Barriere, A. Balahur, Improving sentiment analysis over non-english tweets using multilingual transformers and automatic translation for data-augmentation, arXiv preprint arXiv:2010.03486 (2020).
[24] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. A. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, Advances in Neural Information Processing Systems 35 (2022) 1950–1965.
[25] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Scientific Reports 14 (2024) 6086.
[26] B. Zhang, Z. Liu, C. Cherry, O. Firat, When scaling meets LLM finetuning: The effect of data, model and finetuning method, arXiv preprint arXiv:2402.17193 (2024). The Twelfth International Conference on Learning Representations.
[27] G. H. Chen, S. Chen, Z. Liu, F. Jiang, B. Wang, Humans or LLMs as the judge? A study on judgement biases, arXiv preprint arXiv:2402.10669 (2024).
[28] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou, et al., Larger language models do in-context learning differently, arXiv preprint arXiv:2303.03846 (2023).