The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models

Xiliang Zhu, Shayna Gardiner, Tere Roldán, David Rossouw
Dialpad Inc.
{xzhu, sgardiner, tere.roldan, davidr}@dialpad.com

Abstract

Sentiment analysis serves as a pivotal component in Natural Language Processing (NLP). Advancements in multilingual pre-trained models such as XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) have contributed to the increasing interest in cross-lingual sentiment analysis. The recent emergence in Large Language Models (LLM) has significantly advanced general NLP tasks, however, the capability of such LLMs in cross-lingual sentiment analysis has not been fully studied. This work undertakes an empirical analysis to compare the cross-lingual transfer capability of public Small Multilingual Language Models (SMLM) like XLM-R, against English-centric LLMs such as Llama-3 (AI@Meta, 2024), in the context of sentiment analysis across English, Spanish, French and Chinese. Our findings reveal that among public models, SMLMs exhibit superior zero-shot cross-lingual performance relative to LLMs. However, in few-shot cross-lingual settings, public LLMs demonstrate an enhanced adaptive potential. In addition, we observe that proprietary GPT-3.5 [1] and GPT-4 (OpenAI et al., 2024) lead in zero-shot cross-lingual capability, but are outpaced by public models in few-shot scenarios.

1 Introduction

Sentiment analysis has received considerable attention over the years in the field of Natural Language Processing (NLP) due to its profound value in both academic research and industry applications. Traditionally, studies in sentiment analysis had been mostly focused on high-resource languages such as English due to a deficit of annotated data in other low-resource languages, but recent research has emerged to address this issue by leveraging machine translation to augment data resources (Araújo et al., 2020) (Joshi et al., 2020).

Besides the research efforts in producing multilingual datasets for sentiment analysis, multilingual model architectures have become increasingly popular since the introduction of multilingual pre-trained language models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) and BLOOM (BigScience Workshop, 2022). Such multilingual pre-trained language models exploit the power of large-scale unsupervised textual data from a mixture of many languages, facilitating zero-shot and few-shot cross-lingual transfer from a source to a target language on different downstream NLP tasks, albeit with varying performance outcomes (Lauscher et al., 2020).

More recently, Large Language Models (LLM) such as GPT-3 (Brown et al., 2020), Llama-2 (Touvron et al., 2023) and Llama-3 (AI@Meta, 2024) have collected immense attention for their unparalleled performance in text generation. (Zhang et al., 2023) shows the strong capability of LLMs with few-shot in-context learning in public English sentiment analysis tasks. Although most of the LLMs are pre-trained using corpora with a dominant presence of English, some research has found interesting multilinguality in both public and proprietary LLMs (Qin et al., 2024) (Zhu et al., 2023). Despite these developments, to the best of our knowledge, the capability of cross-lingual transfer in these LLMs has not been fully studied for sentiment analysis tasks, and it is still unclear how LLMs stand in comparison to existing multilingual pre-trained models in the cross-lingual transfer paradigm.

In this work, we examine a variety of pre-trained models and conduct a comprehensive study on the cross-lingual transfer capability in utterance-level sentiment analysis tasks with human speech transcript. We classify our candidate public pre-trained models into two categories: Small Multilin-

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 141–152

gual Language Models (**SMLM**)[2] such as XLM-R and mT5, and more recent Large Language Models (**LLM**)[3] primarily focused on English such as Llama-3 (AI@Meta, 2024) and Mistral (Jiang et al., 2023). In addition, we also include benchmarking with proprietary LLMs such as GPT-4 (OpenAI et al., 2024), which is widely considered as the best LLM in terms of general capability. To avoid potential data contamination introduced in the pre-training process of recent LLMs (Sainz et al., 2023), we curate and annotate proprietary sentiment datasets from in-house human conversation transcripts, and assess cross-lingual sentiment analysis from English to three target languages: Spanish, French and Chinese. Our evaluation results show that with the same supervised fine-tuning, SMLMs demonstrate superior zero-shot cross-lingual transfer capability even with much fewer model parameters. However, public LLMs exhibit rapid improvement in few-shot cross-lingual transfer scenarios and can surpass the performance of SMLMs when additional samples in the target language are provided. Our contributions of this research can be summarized in the following dimensions:

1. We provide a comprehensive comparison on fine-tuning-based cross-lingual transfer capability across a spectrum of public pre-trained language models, with up to 8 billion parameters in the sentiment analysis task on three human languages.

2. Our empirical findings show that some SMLMs (XLM-R, mT5) beat much larger public LLMs in zero-shot cross-lingual transfer. Nevertheless, larger LLMs surpass SMLMs and demonstrate stronger adaptation capability with few-shot fine-tuning in the target language. The best-performing SMLMs still show comparable performance to LLMs when more samples from the target language are provided.

3. We demonstrate that although proprietary GPT-3.5 and GPT-4 present the strongest performance in zero-shot cross-lingual sentiment analysis, with supervised fine-tuning, several public pre-trained language models can outperform GPT-3.5 and GPT-4 in sentiment analysis tasks with few-shot cross-lingual transfer.

[2]We select SMLMs with fewer than 4B parameters in this work.

[3]We select LLMs with at least 7B parameters in this work.

2 Background

2.1 Cross-lingual Sentiment Analysis

Sentiment analysis, as an important subfield of Natural Language Processing, concentrates on detecting and categorizing emotions and opinions in the text. Although the research predominantly focused on the English language initially, subsequent efforts have expanded to support cross-lingual sentiment analysis. This approach aims at leveraging one or several linguistically-rich source languages to enhance task performance in low-resource languages (Xu et al., 2022). Early methods such as (Shanahan et al., 2005) used Machine Translation for cross-lingual sentiment analysis, which became the mainstream methodology in the following years. Other studies focused on bridging the dataset disparities between source and target languages (Zhang et al., 2016), as well as generating parallel corpora for sentiment analysis tasks (Lu et al., 2011) (Meng et al., 2012).

The success of pre-trained models like BERT (Devlin et al., 2019) has spurred adaptations for multilingual and cross-lingual applications, notably mBERT and XLM-R, which utilize a transformer encoder architecture and demonstrate strong capability in cross-lingual language understanding. These models are pre-trained with extensive multilingual corpora and subsequently fine-tuned for specific downstream tasks, thereby significantly enhancing sentiment analysis tasks across diverse languages (Barbieri et al., 2022). (Xue et al., 2021) introduced mT5, which features a transformer encoder-decoder architecture and is pre-trained across over 101 languages, has shown superior performance in classification tasks such as XNLI (Conneau et al., 2018) and surpassed both mBERT and XLM-R. More recently, advancements in unsupervised corpora and computational resources have facilitated the emergence of LLMs with a transformer decoder-only architecture, which have exhibited exceptional performance in various NLP tasks (Touvron et al., 2023) (Jiang et al., 2023) (Brown et al., 2020). Despite these advancements, such LLMs are predominantly English-centric, and their multilingual capabilities remain somewhat ambiguous due to limited disclosure of training data specifics. Furthermore, the capabilities of cross-lingual transfer in these LLMs have yet to be thoroughly studied.

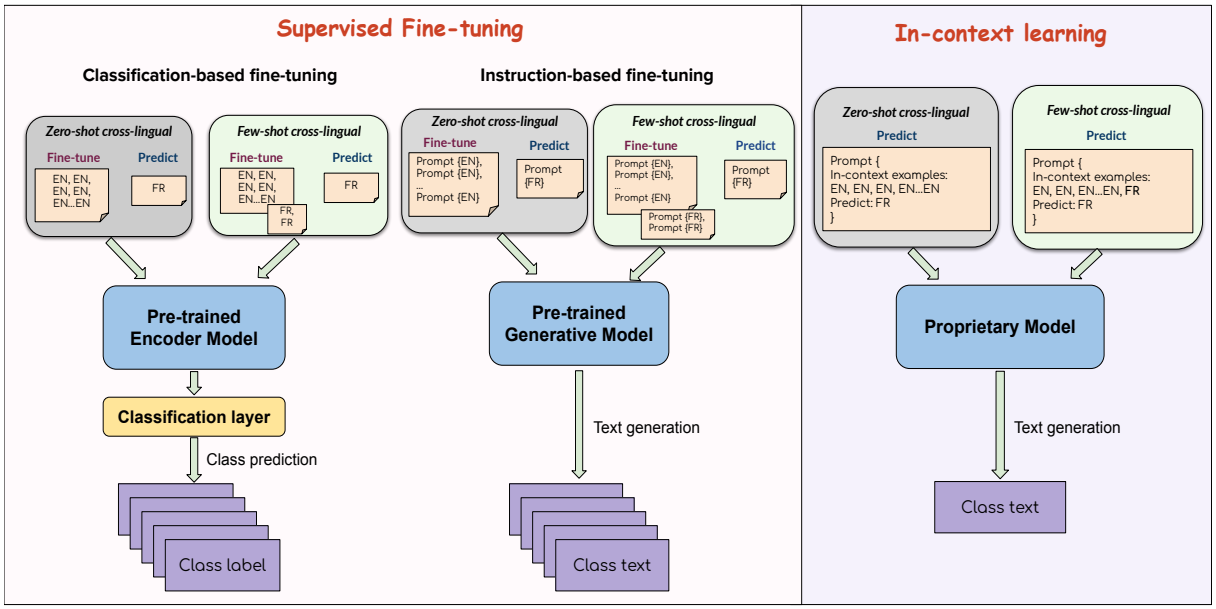

Figure 1: Diagram of zero- and few- shot cross-lingual sentiment analysis from English (EN) to French (FR) under Supervised Fine-tuning (left) and In-context learning (right).

2.2 Sentiment Analysis in Conversational Transcripts

Our work is situated within the context of human conversational transcript data; in our case, these transcript data are obtained from our internal company call centers, consisting of human-to-human conversations that mainly occur between a customer and a customer support agent.

Analyzing such transcript data can be challenging to work with, even for English NLP models: conversational data contain mainly artifacts of spoken language, such as filler words, dysfluencies, and transcription errors by the automated speech recognition (ASR) model (Fu et al., 2022). Adding additional complexity by moving away from English-only data into other languages provides an opportunity to further test the limits of pre-trained language models: switching from one language to another does not always lend itself to a simple, one-to-one translation of each word – especially in describing or expressing abstract concepts like sentiment.

This complexity in cross-lingual sentiment analysis also comes from the need of considering both cultural and linguistic differences. For instance, one of our main observations on sentiment classification in real human conversation in Spanish was that Spanish speakers seem to focus on describing their complaint or situation instead of directly expressing their emotions. For example, they would rather say "*Esta es la quinta vez que los llamo*" ("*This is the fifth time I'm calling you guys*") instead of speaking up and expressing how frustrated they are with a simple and straightforward adjective, such as "*Estoy frustrado*" ("*I am frustrated*"). Whereas the statistical models will easily detect "*frustrado*" and label it as negative sentiment, the abstract description that the speaker chooses in order to express their frustration in the first example will still present a challenge.

3 Methodology

3.1 Supervised Fine-tuning

The objective of this work is to explore the cross-lingual transfer capability of pre-trained models within the context of a sentiment analysis task. To this end, we employ Supervised Fine-tuning (**SFT**) on publicly available pre-trained models using annotated proprietary sentiment datasets (detailed in Section 4.1). Each model is fine-tuned to categorize sentiments as Positive, Negative, or Neutral based on the input provided. Given the diversity in pre-training objectives among different models, we implement two distinct fine-tuning approaches illustrated in Figure 1, which are tailored to the architecture of the pre-trained models:

- **Classification-based fine-tuning**: applicable to transformer encoder-only models such as mBERT and XLM-R, we add a classification layer on top of the pre-trained models and fine-tune the model to directly predict a sentiment

	English (EN)	Spanish (ES)	French (FR)	Chinese (ZH)
Neutral	We're busy, we can't complain, we're fine.	Estamos ocupados, no podemos quejarnos, estamos bien.	Nous sommes occupés, nous ne pouvons pas nous plaindre, nous allons bien.	我们很忙，我们没什么要抱怨的，没事。
	There, I don't know why.	Ahí, no sé por qué.	Là, je ne sais pas pourquoi.	这个，我不知道为什么。
Positive	I love the first one so I'm excited for this one, thanks.	Me encanta el primero, así que estoy emocionado por este, gracias.	J'adore le premier alors je suis excité pour celui-ci, merci.	我很喜欢第一个，对此我感到很兴奋，谢谢。
	This is great, so professional, I'm sure the client was very impressed.	Esto es genial, muy profesional, estoy seguro de que el cliente quedó muy impresionado.	C'est génial, tellement professionnel, je suis sûr que le client était très impressionné.	很好这非常专业，我相信客户一定印象非常深刻。
Negative	I think he's really pissed at me today.	Creo que hoy está muy enojado conmigo.	Je pense qu'il est vraiment très énervé contre moi aujourd'hui.	我感觉他今天对我一定非常生气。
	Yes but I'm worried about being charged twice now.	Sí, pero ahora me preocupa que me cobren dos veces.	Oui mais je suis inquiet d'être facturé deux fois maintenant.	是的，但我对于被收两次费用感到很担心。

Table 1: Examples of our proprietary sentiment datasets.

class.

- **Instruction-based fine-tuning**: used for transformer encoder-decoder (e.g. mT5) and decoder-only (e.g. Llama-3) structures, we construct an instruction to prompt the model to generate a text output corresponding to a sentiment class. The specific prompt format is detailed in Appendix A.1.

To comprehensively evaluate the cross-lingual transfer capabilities of these pre-trained models through fine-tuning, we target both zero- and few-shot cross-lingual transfer from a source to a target language. In *Zero-shot Cross-lingual Transfer* setting, the model is fine-tuned exclusively with an annotated dataset in the source language and subsequently tasked with making predictions in a target language. Note that for generative tasks, merely input language alteration is applied while the instruction component remains constant. *Few-shot Cross-lingual Transfer* extends the zero-shot framework by additionally incorporating N labeled examples from the target language into the fine-tuning process, alongside the source language dataset. The format of the prompt used remains consistent with zero-shot for generative tasks, detailed in Appendix A.1.

3.2 In-context Learning

Recent advancements have highlighted in-context learning as a viable alternative to the traditional fine-tuning approach for generative models (Dong et al., 2023). Due to the access limitation and our data privacy policy, we are not able to fine-tune proprietary LLMs using our proprietary datasets. Consequently, we employ in-context learning through the prompt to simulate an experiment setting as conducting SFT on public models. Nonetheless, the inherent limitation regarding the context length in various close source LLMs poses a challenge; these models may not accommodate as many examples within a prompt as is feasible for SFT in open source counterparts. Figure 1 shows an illustrative diagram of in-context learning for this sentiment analysis task.

To assess cross-lingual transfer capabilities as Section 3.1 through in-context learning, we construct in-context examples with different sources of languages accordingly. Specifically, for *Zero-shot Cross-lingual Transfer*, the prompts include examples solely from the source language. In contrast, for *Few-shot Cross-lingual Transfer*, additional supplementary examples in the target language are also applied. Prompts with in-context examples we use to evaluate proprietary LLMs are attached in Appendix A.2.

4 Experiment

In this section, we first present a detailed description of our internal proprietary sentiment datasets which are used for fine-tuning and evaluation. Then, we provide necessary introductions to a diverse array of public pre-trained models we will study for this work. Finally, we show the hardware and software resources employed in conducting the experiment.

Model type	Name	Architecture	# of param.	Claimed language support
SMLM	mBERT	encoder	110M	104 langs
	XLM-R-base	encoder	250M	100 langs
	XLM-R-large	encoder	560M	100 langs
	mT5-base	encoder-decoder	580M	101 langs
	mT5-large	encoder-decoder	1.2B	101 langs
	mT5-xl	encoder-decoder	3.7B	101 langs
English-centric LLM	Mistral-7B	decoder	7B	Unclear
	Falcon-7B	decoder	7B	Mainly EN, DE, ES, FR
	Llama2-7B	decoder	7B	Intended for EN
	Llama3-8B	decoder	8B	Intended for EN

Table 2: List of public pre-trained models evaluated in our experiments.

4.1 Dataset

The proprietary datasets used in this study are utterance-level sentiment data for four languages: English, Spanish, French, Chinese (Table 1). Utterance boundaries are generated by our in-house ASR system when a short pause or speaker change is detected in the audio stream. We randomly sampled English and Spanish utterances from the real conversational transcript from our call center applications and each instance is labeled as **Positive**, **Negative** or **Neutral** by human annotators. The annotation was done via a third-party vendor, allowing us to configure our ontology and direct the annotators to select the appropriate category for the sentiment detected in each utterance according to guidelines we developed. Our guidelines include definitions for each sentiment as well as a broad list of examples (a gold dataset manually annotated by our internal team). Inter-annotator agreement is calculated automatically by our annotation vendor, and a high agreement threshold is applied to ensure the quality of the annotation results.[4]

Constrained by resources, we are not able to sample and annotate French and Chinese datasets under the same setting. Instead, we leverage machine translation (through GPT-4, detailed in Appendix A.3) to create parallel French and Chinese datasets based on the annotated English counterpart. All machine-translated datasets were reviewed by speakers of the target language to ensure that the translations were comparable to the original English. There were some minor issues identified in the machine-translated data during review: namely, occasionally GPT-4 refuses to translate a sample, producing a refusal in the target language instead, or it produced a commentary on the English transcript in the target language in lieu of translating it directly. These samples were identified and removed, and the remaining samples were deemed to be accurate translations by the speakers of the target languages.

As our objective is to study the cross-lingual sentiment analysis from English to target languages, we assemble English data with a much larger size, while Spanish, French and Chinese with a limited amount sufficient only to support few-shot learning and testing purposes. A summary of the total amount of data used for the following experiment is as follows:

- English: 30,000 instances for fine-tuning, 3,000 for development.
- Spanish: 600 instances for fine-tuning and 3,000 for testing.
- French: 600 instances for fine-tuning and 3,000 for testing.
- Chinese: 600 instances for fine-tuning and 3,000 for testing.

where we ensure sentiment labels are uniformly distributed across all sets.

Table 1 shows exemplary cases of our proprietary datasets in different languages, providing insight into domain-specific textual characteristics. It is worth mentioning that these examples have no identifying information and are intended for illustrative purposes only. The use of internal call transcript data ensures that all model evaluations are immune from unintended data contamination of the pre-trained models, which could otherwise lead to an overestimation of their performance (Sainz et al., 2023).

4.2 Selected pre-trained Models

In this work, we investigate a variety of public pre-trained language models, with a range of sizes and architectures. For SMLM, we have selected models from mBERT, XLM-R and mT5 model

[4] https://docs.labelbox.com/docs/consensus

	Public SMLM						Public LLM				Proprietary LLM	
	Supervised Fine-tuning						Supervised Fine-tuning				In-context Learning	
	mBERT	**XLM-R-base**	**XLM-R-large**	**mT5-base**	**mT5-large**	**mT5-xl**	**Mistral**	**Falcon**	**Llama-2**	**Llama-3**	**GPT-3.5**	**GPT-4**
	110M	250M	560M	580M	1.2B	3.7B	7B	7B	7B	8B	-	-
ES	47.1	54.4	58.7	60.2	63.4	60.0	44.8	55.3	60.1	57.9	75.6	74.8
FR	45.3	71.8	76.8	75.4	79.7	73.8	48.4	70.7	74.5	77.4	80.3	79.3
ZH	54.2	72.3	76.9	74.8	77.3	71.5	40.4	71.9	64.9	73.3	82.3	80.2
Avg	48.9	66.2	70.8	70.1	**73.5**	68.4	44.5	66.0	66.5	69.5	**79.4**	**78.1**

Table 3: F1 score comparison in zero-shot cross-lingual transfer on our proprietary sentiment analysis datasets. ES: Spanish, FR: French, ZH: Chinese. Top-3 average F1 scores are marked in bold.

Figure 2: Average F1 score performance comparison (across ES, FR and ZH) under N-shot settings. GPT-3.5 is not included in this 600-shot due to the context length limit.

families with up to 3.7 billion parameters. All models in our SMLM selection are known for their support for over 100 human languages and have demonstrated efficacy in tasks that require multilingual and cross-lingual capabilities, as evidenced by references (Doddapaneni et al., 2021) (Xue et al., 2021). For English-centric LLMs, the details are little disclosed regarding the specific human languages incorporated during the pre-training phase. Therefore, we include the most prominent and widely recognized models from Llama family and Mistral with 7 to 8 billion parameters sizes. In additional, Falcon-7B is also added to our analysis as it explicitly claims proficiency in German, Spanish and French in addition to English. The specifics of all the pre-trained models utilized in our experiments are detailed in Table 2.

4.3 Experiment Setup

The fine-tuning and inference processes for our model are conducted using the Huggingface framework (Wolf et al., 2020) on a single-node Linux system equipped with eight Nvidia A100 80G GPUs. For experiments on proprietary LLMs, we use "`gpt-3.5-turbo-0125`" endpoint for GPT-3.5 and "`gpt-4-1106-preview`" endpoint for GPT-4.

In order to ensure deterministic output from generative models, temperature is set as 0 for all public and proprietary models in our experiments.

5 Results

To facilitate a comprehensive comparison between SMLMs and LLMs on cross-lingual sentiment analysis, we follow the zero-shot and few-shot cross-lingual fine-tuning methodologies described in 3.1 and evaluate the model performance respectively. The F1 score (micro) is employed as the accuracy evaluation metric in the following sentiment analysis experiments.

5.1 Zero-shot Cross-lingual Transfer

We first fine-tune public pre-trained models in zero-shot cross-lingual transfer setting through SFT as detailed in Section 3.1, exposed to only the English

fine-tuning dataset described in 4.1. Note that we leverage in-context learning for proprietary LLMs as discussed in Section 3.2. However, due to constraints on context length, these proprietary LLMs are not exposed to the entirety of the English fine-tuning set; instead, they are prompted with a set of 300 examples, carefully balanced across different classes for this experiment.

Evaluation results are presented in Table 3. It is clear that both GPT-3.5 and GPT-4 exhibit significant advantages over fine-tuned public models on target languages in zero-shot. Surprisingly, among the public models, several SMLMs such as XLM-R-large (560M), mT5-base (580M) and mT5-large (1.2B), show better zero-shot cross-lingual transfer capability compared to the considerably larger Mistral-7B, Falcon-7B, Llama2-7B and Llama3-7B models. In particular, mT5-large surpasses all other open source candidates by a substantial margin across all testing languages despite having only 1.2 billion parameters.

5.2 Few-shot Cross-lingual Transfer

We then fine-tune and evaluate public models under the few-shot cross-lingual transfer setting described in Section 3, where we randomly select ***N*** training samples in the target language and use them in fine-tuning in conjunction with the English fine-tuning data. In order to better investigate the adaptability of the models, we vary ***N*** among {60, 150, 600}, thereby conducting **60-shot**, **150-shot** and **600-shot** experiments respectively. The selection of these three values provides a wide spectrum for comparative analysis, also ensures a sufficient representation while maintaining resource-efficient. For proprietary LLMs, an additional ***N*** samples in target language are appended to the prompt during in-context learning to establish a similar few-shot cross-lingual setup.

The evaluation results of average F1 scores across three target languages (ES, FR and ZH) are presented in Figure 2, under the settings of 60-shot, 150-shot and 600-shot. Detailed F1 scores per language are also provided in Appendix A.4. Our observations and findings can be summarized as follows:

i Among public pre-trained models, despite their underperformance relative to SMLMs in zero-shot cross-lingual transfer as evidenced in Table 3, English-centric LLMs present strong adaptation capability in few-shot cross-lingual sentiment analysis. Notably, all public LLMs exhibit significant relative improvements compared to their zero-shot performance. It is worth pointing out that with 60-shot and 150-shot, LLMs such as Falcon-7B, Llama2-7B and Llama3-8B surpass the performance of all SMLMs by a considerable margin. The only exception is Mistral-7B, which is still outperformed by several SMLMs with few-shot.

ii With an increased volume of training data in the target language, specifically under 600-shot condition, mT5-xl with 3.7B parameters has a comparable performance to the much larger Falcon-7B, Llama2-7B and Llama3-8B models.

iii Contrary to their dominance in the zero-shot cross-lingual setting, GPT-4 and GPT-3.5 exhibit very limited improvement in few-shot cross-lingual sentiment analysis with in-context examples. Several public models are capable of surpassing these prominent proprietary LLMs following fine-tuning.

6 Conclusion

In this study, we explore the capabilities of cross-lingual sentiment analysis across a variety of pre-trained language models. We show that smaller XLM-R-large (560M), mT5-base (580M) and mT5-large (1.2B) have superior zero-shot cross-lingual transfer capabilities compared to the considerably larger Mistral-7B, Falcon-7B, Llama2-7B and Llama3-8B models. This highlights the efficiency and potential of Small Multilingual Language Models (SMLM) for sentiment analysis in low-resource languages. On the other hand, our findings reveal that the larger English-centric LLMs like Falcon-7B and Llama2-7B can quickly adapt and show much improved performance with a few-shot cross-lingual setup, which indicates their robustness in learning from limited data from the target language. Moreover, proprietary LLMs such as GPT-3.5 and GPT-4 exhibit the strongest zero-shot performance in cross-lingual sentiment analysis tasks, however, in scenarios involving few-shot learning, several fine-tuned public pre-trained models are able to surpass these proprietary giants.

7 Limitation

Although our findings in this study appear to be consistent in all target languages tested, due

to the limitation of our resources, it is still unclear how the models would behave in other low-resource languages with even less appearance during pre-training. In addition, due to the incomparable model sizes, we are not able to draw any conclusions on whether model architecture difference (transformer encoder-only, decoder-only and encoder-decoder) could play a role in cross-lingual sentiment analysis capabilities. Further research could be extended in these directions.

References

AI@Meta. 2024. Llama 3 model card.

Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

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.

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models. *Preprint*, arXiv:2107.00676.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Xue-yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shayna Gardiner, Pooja Hiranandani, and Shashi Bhushan Tn. 2022. Entity-level sentiment analysis in contact center telephone conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 484–491, Abu Dhabi, UAE. 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. *Preprint*, arXiv:2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 320–330, Portland, Oregon, USA. Association for Computational Linguistics.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual

mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–581, Jeju Island, Korea. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

James Shanahan, Gregory Grefenstette, Yan Qu, and David Evans. 2005. Mining multilingual opinions through classification and translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten,

Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Yuemei Xu, Han Cao, Wanze Du, and Wenqing Wang. 2022. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Science and Engineering*, 7:1–21.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Peng Zhang, Suge Wang, and Deyu Li. 2016. Cross-lingual sentiment classification: Similarity discovery plus training data adjustment. *Knowledge-Based Systems*, 107:129–141.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A Appendix

A.1 Prompt Format for Supervised Fine-tuning

We employ the following prompt format in supervised fine-tuning for public generative models:

```
Below is an utterance extracted from the transcript of a business call, identify the speaker's sentiment in this utterance. The sentiment should be one of the following:
"Positive": The speaker expresses favorable emotions and mental states, for example, euphoria and joy, happiness, excitement, fascination, satisfaction, pride, gratitude, relief, surprise, etc.
"Negative": The speaker expresses unfavorable emotions and mental states, for example, disgust, sadness, disappointment, worry, insecurity, annoyance, fury, anger, fear, depression, frustration, etc.
"Neutral": Statement in which the speaker does not express emotions, but in which a fact is simply stated and no explicit emotions or feelings are conveyed.
What is the sentiment in the following utterance? Only respond with the sentiment without explanation:
### Input: {utterance text}
### Output:
```

A.2 Prompt Format for In-context Learning

The following prompt with in-context examples is used for calling proprietary LLM APIs:

```
Below is an utterance extracted from the transcript of a business call, identify the speaker's sentiment in this utterance. The sentiment should be one of the following:
"Positive": The speaker expresses favorable emotions and mental states, for example, euphoria and joy, happiness, excitement, fascination, satisfaction, pride, gratitude, relief, surprise, etc.
"Negative": The speaker expresses unfavorable emotions and mental states, for example, disgust, sadness, disappointment, worry, insecurity, annoyance, fury, anger, fear, depression, frustration, etc.
"Neutral": Statement in which the speaker does not express emotions, but in which a fact is simply stated and no explicit emotions or feelings are conveyed.
Here are some examples:
### Input: {utterance text 1}
### Output: {sentiment label 1}
```

```
### Input:  {utterance text 2}
### Output:  {sentiment label 2}

### Input:  {utterance text 3}
### Output:  {sentiment label 3}

...

What is the sentiment in the follow-
ing utterance?  Only respond with the
sentiment without explanation:
### Input:  {utterance text}
### Output:
```

A.3 Machine translation details

The machine translation process described in Section 4.1 utilizes GPT-4 endpoint "`gpt-4-1106-preview`". The prompt used for machine translation is as follows:

```
Below is a transcribed utterance from
human conversations, translate it from
English to {TARGET_LANG}:
### Input:  {English utterance}
### Output:
```

TARGET_LANG refers to the target languages in our machine translation process, i.e. French and Chinese.

A.4 Per-language Evaluation Tables for Few-shot Cross-lingual

Supplementary to Section 5.2, detailed per language evaluation results on few-shot cross-lingual are listed in Table 4, Table 5, and Table 6

	Public SMLM						Public LLM				Proprietary LLM	
	Supervised Fine-tuning						Supervised Fine-tuning				In-context Learning	
	mBERT	**XLM-R-base**	**XLM-R-large**	**mT5-base**	**mT5-large**	**mT5-xl**	**Mistral**	**Falcon**	**Llama-2**	**Llama-3**	**GPT-3.5**	**GPT-4**
	110M	250M	560M	580M	1.2B	3.7B	7B	7B	7B	8B	-	-
ES	71.0	62.7	67.1	59.7	65.3	73.2	73.1	76.8	77.7	77.6	76.0	76.8
FR	69.3	79.7	82.7	76.1	83.7	83.8	76.1	82.3	84.7	85.2	81.6	80.3
ZH	73.7	80.0	81.7	78.0	80.8	80.7	74.9	84.0	81.2	83.5	80.1	80.4
Avg	71.3	74.1	77.2	71.3	76.6	79.2	74.7	**81.0**	**81.2**	**82.1**	79.2	79.2

Table 4: F1 score comparison in **60-shot** cross-lingual transfer on our proprietary sentiment analysis datasets. ES: Spanish, FR: French, ZH: Chinese. Top-3 average F1 scores are marked in bold.

	Public SMLM						Public LLM				Proprietary LLM	
	Supervised Fine-tuning						Supervised Fine-tuning				In-context Learning	
	mBERT	**XLM-R-base**	**XLM-R-large**	**mT5-base**	**mT5-large**	**mT5-xl**	**Mistral**	**Falcon**	**Llama-2**	**Llama-3**	**GPT-3.5**	**GPT-4**
	110M	250M	560M	580M	1.2B	3.7B	7B	7B	7B	8B	-	-
ES	71.9	71.6	71.8	60.5	69.4	74.7	71.1	76.8	79.7	77.6	76.3	74.5
FR	71.3	82.0	82.9	78.0	83.3	83.0	76.0	86.1	84.2	82.9	81.9	78.7
ZH	76.8	82.7	84.1	78.4	81.7	83.6	78.7	84.5	85.6	85.2	81.7	82.6
Avg	73.3	78.8	79.6	72.3	78.1	80.4	75.3	**82.5**	**83.2**	**81.9**	80.0	78.6

Table 5: F1 score comparison in **150-shot** cross-lingual transfer on our proprietary sentiment analysis datasets. ES: Spanish, FR: French, ZH: Chinese. Top-3 average F1 scores are marked in bold.

	Public SMLM						Public LLM				Proprietary LLM	
	Supervised Fine-tuning						Supervised Fine-tuning				In-context Learning	
	mBERT	**XLM-R-base**	**XLM-R-large**	**mT5-base**	**mT5-large**	**mT5-xl**	**Mistral**	**Falcon**	**Llama-2**	**Llama-3**	**GPT-3.5**	**GPT-4**
	110M	250M	560M	580M	1.2B	3.7B	7B	7B	7B	8B	-	-
ES	74.0	74.0	77.4	64.4	77.9	77.6	76.3	79.0	79.0	76.6	-	73.9
FR	76.1	83.7	83.8	79.9	86.2	87.4	83.6	86.6	86.8	86.1	-	78.8
ZH	81.8	85.8	86.4	80.9	83.8	88.6	87.8	88.3	88.0	89.3	-	81.4
Avg	77.3	81.2	82.5	75.1	82.6	**84.5**	82.6	**84.7**	**84.6**	84.0	-	78.0

Table 6: F1 score comparison in **600-shot** cross-lingual transfer on our proprietary sentiment analysis datasets. ES: Spanish, FR: French, ZH: Chinese. Top-3 average F1 scores are marked in bold. GPT-3.5 is not included in this evaluation due to the context length limit.