Effectiveness of Scalable Monolingual Data and Trigger Words Prompting on Cross-Lingual Emotion Detection Task

Yao-Fei Cheng*, **Jeongyeob Hong***, **Andrew Wang***, **Anita Silva***, **Gina-Anne Levow**
Department of Linguistics
University of Washington
{nlp5566,yeob,andrewzw,silvaa5,levow}@uw.edu

Abstract

This paper introduces our submitted systems for WASSA 2024 Shared Task 2: Cross-Lingual Emotion Detection. We implemented a BERT-based classifier and an in-context learning-based system. Our best-performing model, using English Chain of Thought prompts with trigger words, reached 3rd overall with an F1 score of 0.6015. Following the motivation of the shared task, we further analyzed the scalability and transferability of the monolingual English dataset on cross-lingual tasks. Our analysis demonstrates the importance of data quality over quantity. We also found that augmented multilingual data does not necessarily perform better than English monolingual data in cross-lingual tasks. We open-sourced the augmented data and source code of our system for future research.[1]

1 Introduction

Recognizing the affect of tweets presents a crucial challenge as they encapsulate the semantic and emotional dialogue of individuals spanning diverse cultures and languages through a short, informal, and noisy medium. The unique constraint of limited length allows users to communicate through abbreviations, slang, and emojis. Such lexical idiosyncrasies, compounded by fragmented language and cultural references, further exacerbate the challenge for traditional natural language processing (NLP) models to interpret emotional content in tweets.

Previous studies have explored various approaches to sentiment and emotion classification (Mohammad, 2016). The SemEval-2018 Task 1 (Mohammad et al., 2018) presented an array of tasks for recognizing the affect of tweets, focusing on building monolingual English, Arabic, and Spanish systems. The vast majority of prior research on affect recognition models has been conducted on monolingual English data, but an increasing number of studies are now focused on the use of cross-lingual models to improve emotion classification (De Bruyne, 2023). WASSA 2024 Shared Task 2 follows the trend of multilingual emotion classification with an emphasis on model explainability and interpretability (Maladry et al., 2024). With the advent of large language models (LLM) accessible through closed API calls, explaining the rationale of LLM prediction is increasingly important.

In our participation in the WASSA 2024 Shared Task 2, we present key observations in data selection and monolingual vs. multilingual approaches through our comparison of our BERT-based models to LLM prompting methods. Our findings indicate that pre-training with in-domain data yields better performance than within-task data. Additionally, monolingual data (back-translated) outperforms multilingual data. We also introduce our balanced dataset for fine-tuning and make it open-source for future research[2].

2 System Overview

2.1 BERT-based classifier

We summarize our system in Figure 1, and detailed hyper-parameters can be found in Table 5. The system consists of 1) data augmentation, 2) pre-processing, 3) creating our Treehouse LM by continued pre-training the pre-trained TwHIN-BERT-Large (Zhang et al., 2022) language model with augmented data, and 4) fine-tuning the Treehouse LM with in-domain tweets for the downstream task.

2.1.1 Data Augmentation

We observed a significant label imbalance for the provided official English training set (Official Train), as seen in Figure 3. To boost the robustness of our system, we used the English SemEval 2018

[1] * Equal contribution.

[2] https://github.com/freddy5566/cross-lingual-emotion-detection

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 511–522

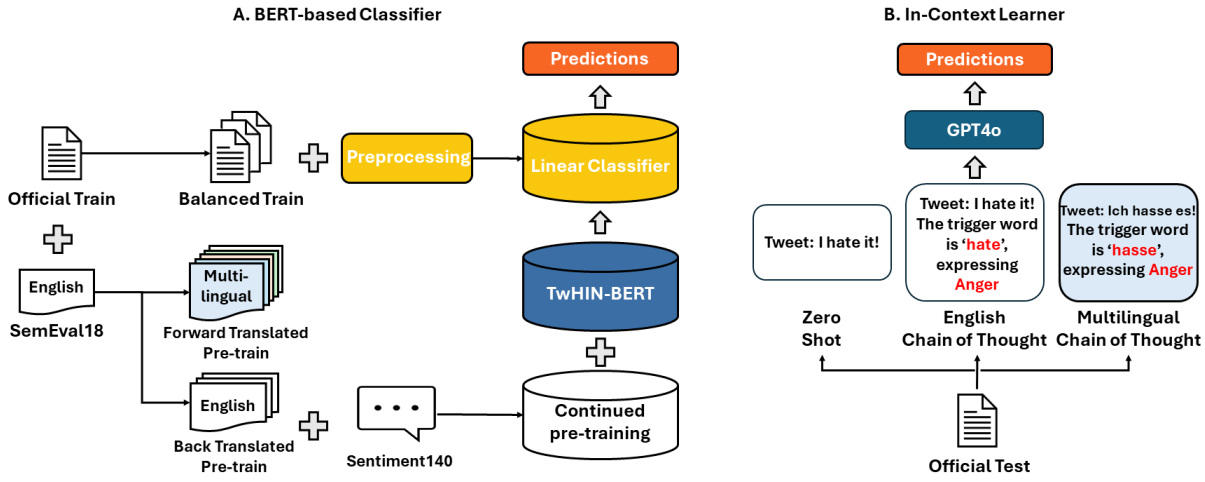

Figure 1: The left figure (A) shows the workflow of the BERT-based classifier, indicating the sources of augmented data for fine-tuning and continued pre-training. After continued pre-training, the model is fine-tuned to produce predictions. The right figure (B) illustrates our prompt-based system, which uses official test data and adds Chain of Thought examples, sending it directly to GPT4o for predictions.

Task1 dataset (Mohammad et al., 2018), and Sentiment140 (Go et al., 2009) to create an augmented training set for both fine-tuning and continued pre-training.

To build a balanced train set (Balanced Train) for fine-tuning, we filtered the English SemEval 2018 Task 1 dataset for entries that were labeled with one of the five underrepresented labels: Joy, Anger, Love, Sadness, or Fear. Then, we randomly selected the samples, reaching $6,970$ entries.

To build the augmented train set for continued pre-training (Back Translated Pre-train), we started with translating Arabic and Spanish samples in the SemEval 2018 Task 1 dataset into English. Combined with the Official Train dataset, the dataset reached a size of $27,458$. It was further back-translated with four target languages: Spanish, Russian, French, and Dutch, yielding a set of $132,290$ entries after removing instances from Official Train to prevent data contamination. All translations are conducted through Google Translation API.

Furthermore, we combined our Back Translated Pre-train dataset with the Sentiment140 dataset[3], a tweet dataset with 1.6 million instances. We experimented with the various combinations of the two datasets to show the scalability and transferability of the monolingual English dataset in cross-lingual emotion recognition. Table 1 summarizes the data used in the system.

To further compare the effect of language variety in training data, we also created a multilingual training set for fine-tuning and continued pre-training. Details are provided in the Appendix A.2. The multilingual dataset was used only for paper-writing purposes and was thus not used during the competition.

Data	# of Sentences
Official Train	5,000
Balanced Train	6,970
Back Translated Pre-train	132,290
+ Half Sentiment140	1,000,000
+ Full Sentiment140	1,737,290
Sentiment140	1,600,000

Table 1: Data Distribution

2.1.2 Pre-processing

Through experimentation, we found that certain pre-processing methods, such as lowercasing, auto-correction, and slang replacement, degraded performance. The impact of removing hashtags, numbers, and punctuation varied, as these elements can both convey important emotional context and introduce noise.

The optimal pre-processing methods we identified were: converting Unicode symbols to ASCII, removing mentions and links, eliminating repetitive characters, and converting over-segmented character sequences back into words. These choices

[3]https://www.kaggle.com/discussions/product-feedback/176309

	Base	Zeroshot	EngCoT	MulCoT	TwHIN	$130k$	$1M$	$1.7M$	Fwd
F1	0.4476	0.5872	**0.6015**	0.5966	0.5010	0.5220	0.5164	0.5171	0.5163
Precision	0.4452	0.5993	**0.6039**	0.5879	0.4944	0.5204	0.5284	0.5100	0.5225
Recall	0.4631	0.5798	0.6085	**0.6125**	0.5205	0.5311	0.5117	0.5314	0.5247

Table 2: EngCoT and MulCoT denote the English and multilingual prompts-based ICL methods, respectively. TwHIN column presents results without additional continued pre-training. $130k$, $1M$, and $1.7M$ each represent the size of the dataset used in continued pre-training. Fwd refers to the model trained with multilingual data.

indicate that TwHIN-BERT, trained on extensive Twitter data, effectively handles the inherent noise in tweets and that excessive pre-processing can overcorrect testing data and degrade its ability to capture semantic representations.

2.1.3 Continued pre-training

Following the findings in Gururangan et al. (2020) of domain and task-adaptive continued pre-training, we continue to pre-train BERT-based LMs (Devlin et al., 2019; Liu et al., 2019; Nguyen et al., 2020; Conneau et al., 2020; Zhang et al., 2022) to further adapt the general Twitter domain to the specific Twitter domain of EXALT data with the augmented data introduced in Section 2.1.1.

2.1.4 Fine-tuning

We fine-tuned the continued pre-trained LM for downstream emotion classification with the Balanced Train data. We add one linear classifier on top of the pre-trained LM that takes the CLS token x as the input and applies a linear projection $T : \mathbb{R}^d \rightarrow \mathbb{R}^n$, where d is the dimension of CLS token embedding and n is the number of classes. The final distribution can be obtained by applying the softmax function $\sigma(T(x))$.

2.2 In-context Learner

Brown et al., 2020 found that large language models can learn the contextual information from raw text, indicating that they can be used for downstream tasks without fine-tuning on additional labeled data, a technique called in-context learning (ICL). In this section, we describe our method for building an ICL-based system.

2.2.1 Chain of Thought with trigger words

Chain-of-Thought (CoT) is a method that introduces intermediate steps that decompose the reasoning process to the LLMs, enabling them to significantly improve their performance (Wei et al., 2022). To ensure the explainability of affect recognition, we designed English CoT prompts (EngCoT) that explain the classification process by identifying the trigger word. As shown in Figure 1, the given prompt points out that "hate" is the trigger word for the tweet "I hate it." We provided an explanation for the Neutral label where no clear trigger word exists based on the annotation guidelines (Singh, Pranaydeep and Maladry, Aaron and Lefever, Els, 2023). Example trigger words and tweets are selected from the trigger word detection dataset from subtask 2 of EXALT Shared Task.

Our system uses GPT4o, which performed best in comparison to other models. See A.6.2 for the details. Similar to the BERT-based classifier, we further analyzed system performance with multilingual CoT prompts (MulCoT). The experiment details and full prompts used in our system are provided in Appendix A.6.

3 Results

Table 2 summarizes our system results. The overall best-performing model was GPT4o with English Chain of Thought instructions (EngCoT). This system also reached 3rd place overall in the leaderboard. However, the best recall was achieved through multilingual CoT (MulCoT).

Regarding BERT-based classifiers, continued pre-training TwHIN-BERT with 130k back-translated data outperformed 1M and 1.7M models. A detailed ablation study can be found in Appendix A.3.

4 Discussion

4.1 Scalability of continued pre-training

We examine whether data quantity or quality is more important in pre-training. As suggested in (Sun et al., 2019), continued pre-training with in-domain or within-task data can provide a certain level of improvement to the downstream task. We used Back Translated Pre-train as the in-domain data because of the similarity of data distribution with fine-tuning data. We used Sentiment140 as

the within-task pre-training data because it fits in the sentiment classification task to conduct this examination.

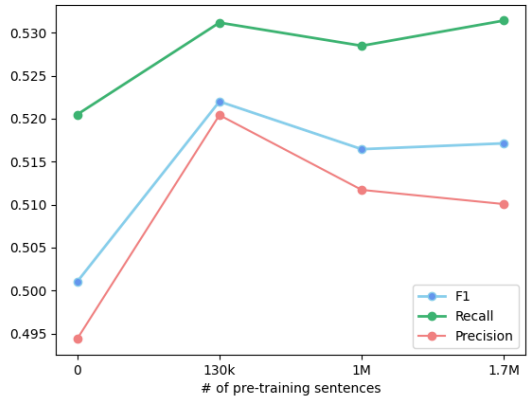

Figure 2: The impact of the scale of pre-training data.

As we can see in Figure 2, all models with continued pre-training perform better than those with no pre-training. Among them, pre-training with in-domain data achieved the best performance. On the other hand, the performance of models (1M and 1.7M) trained with within-task data slightly trails that of the model (130k) trained with in-domain data despite the huge difference in the amount of training data in the pre-training stage. This result suggests that the similarity of data distribution between pre-training and fine-tuning is key to yielding the best performance.

4.2 Multilingual vs. Monolingual

4.2.1 BERT-based classifier

	F1	Recall	Precision
Balanced Train (7k)	0.5010	0.5204	0.4943
Subsampled (7k)	0.4663	0.4908	0.4657
All (25k)	**0.5163**	**0.5225**	**0.5247**
+ Multi. PT (130k)	0.5061	0.5125	0.5107

Table 3: The results on multilingual training data trained with TwHIN-Bert-Large. PT represents seed models with continued pre-training. Subsampled indicates Balanced Translated Train, and All indicates Forward Translated Train.

To investigate the effectiveness of multilingual data, we compared the multilingual model (please refer to Appendix A.2 for a detailed description) and back-translated English-only monolingual model.

The results, shown in Table 3, show our English back-translated model "Balanced Train" performing better than the multilingual model "Subsampled" under the same amount of training data. However, the multilingual model "All" outperforms "Balanced Train". We hypothesize that performance gain in "All" comes from the amount of training data. Furthermore, our experiments demonstrate that, unlike monolingual continued pre-training, multilingual continued pre-training can be harmful. The root causes still need further study.

4.2.2 Prompting-based methods

Similar to our findings with the BERT-based classifier, the use of multilingual CoT did not outperform monolingual English CoT prompts. Further analysis of the output labels from EngCoT and MulCoT showed that both methods improve the F1 scores by enhancing Recall at the expense of Precision (Table 11 and 12). Compared to EngCoT, MulCoT resulted in a notable drop in Precision. This trend is consistent, as on average, over five runs, EngCoT achieved results of 0.5923, 0.5972, and 0.5997 in F1, Recall, and Precision, respectively. In comparison, MulCoT resulted in 0.5893, 0.6046, and 0.5863. The reason for such difference remains unanswered, but we present our additional experiments on prompts in Section A.6.3 for future research.

5 Conclusion

This paper describes our proposed systems for Shared Task 2 of WASSA 2024. Through the analysis of BERT-based classifiers and in-context learning-based systems, we highlight the importance of high-quality data and the effectiveness of CoT using trigger words. In our continued pre-training experiments, we discovered that aligning the data distribution between continued pre-training data and fine-tuning data is crucial. Without this alignment, the size of the dataset does not significantly impact information transferring into cross-lingual settings. While it may seem intuitive to use multilingual data for cross-lingual tasks, our findings revealed that this approach did not enhance performance in both systems. Further research is needed to understand the underlying mechanisms of in-context learning and its impact on performance.

6 Limitations

The main limitations of our work relate to these points: a) Our augmented data is highly dependent on the quality of Google Translation API. Furthermore, it is not a deterministic output; b) We did not perform an extensive hyper-parameter search in continued pre-training, which might improve classifiers' performance. c) We presented outputs from closed-source models, where access is limited through paywall APIs. d) The outputs from GPT-4 are not deterministic and are vulnerable to changes in prompts. e) Considering the similarity between sentiment classification and emotion detection, we treated the sentiment 140 data as within-task data. However, those two tasks are related, not identical; therefore, results could be affected by such differences.

References

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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.

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*.

Luna De Bruyne. 2023. The paradox of multilingual emotion detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*.

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 57th Annual Meeting of the Association for Computational Linguistics*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the wassa 2024 exalt shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop of on Computational Approaches to Subjectivity, Sentiment Social Media Analysis@ACL 2024*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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*.

Saif M. Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herbert L. Meiselman, editor, *Emotion Measurement*, pages 201–237. Woodhead Publishing.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Singh, Pranaydeep and Maladry, Aaron and Lefever, Els. 2023. Annotation guidelines for labeling emotion in multilingual tweets.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and

Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of 36th Conference on Neural Information Processing Systems*.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. *arXiv preprint arXiv:2209.07562*.

A Appendix

A.1 Data

Figure 3 shows the label distribution in Official and Balanced Train.

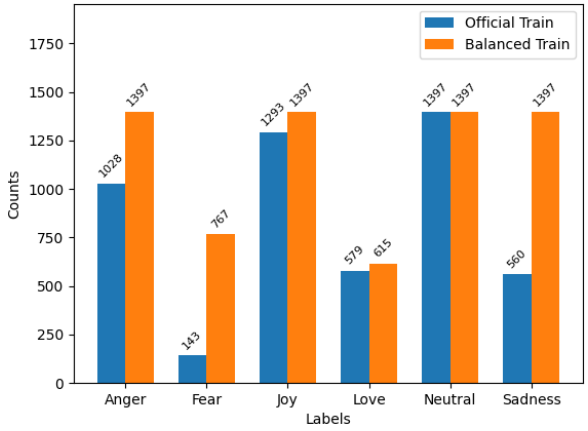

Figure 3: Comparison of Official train set with enhanced train set.

A.2 Multilingual Training Data

Data	# of Sentences
Official Train	5,000
Balanced Train	6,970
Back Translated Pre-train	132,290
Forward Translated Train	25,000
Balanced Translated Train	6,923
Forward Translated Pre-train	134,745

Table 4: Data Distribution

This section explains the process of building multilingual training data for fine-tuning and continued pre-training. Likewise, we translated the Official Train into four other languages and combined it with the Official Train, resulting in $25,000$ entries for the multilingual training set (Forward Translated Train). To have a fair comparison with Balanced Train, we also subsample 6,923 entries with a similar label distribution to Balanced Translated Train.

For the multilingual continued pre-training dataset (Forward Translated Pre-train), we combined $27,458$ English data, created by translating Arabic and Spanish SemEval data in English, and translated it into French, Dutch, Spanish, and Russian. After removing identical entries, the Forward Translated Pre-train contains $134,745$ entries. Table 4 summarizes the data.

A.3 Classifiers

In this section, we conducted a detailed ablation study on BERT-based classifiers. We examined five strong pre-trained LMs (Devlin et al., 2019; Liu et al., 2019; Nguyen et al., 2020; Conneau et al., 2020; Zhang et al., 2022) with different conditions. The detailed hyper-parameters are summarized in Table 5.

Hyper-parameter	Pre-training	Fine-tuning
Max sequence length	128	512
Precision	FP16	FP16
Total batch size	128	64
Learning rate	1e-4	2e-6
# of epoch	10	20
Weight decay	0.1	0.1

Table 5: The hyper-parameters for pre-training and fine-tuning.

A.3.1 Which pre-trained LM as the seed model?

Our results are summarized in Table 6. We found TwHIN-BERT performed best regardless of whether it was pre-processed or not and whether the test set was translated into English. Therefore, we use TwHIN-BERT as our seed model in the rest of the experiments.

A.3.2 Pre-processing

Surprisingly, models without pre-processing generally perform better than those with pre-processing. We hypothesize that this is because TwHIN-BERT is capable of capturing and representing the deep lexical and semantic properties of tweets, having been trained on complete Twitter data. Although noisy, various lexical features of tweets are strong indicators of emotional context, unique tokens such as emoticons and emojis can be challenging for the model to tokenize and represent accurately. Therefore, a promising area for future research is to explore methods for converting these tokens into meaningful words or alternative representations.

A.4 English dev/test

Another obvious trend is translating dev and test sets into English, dramatically improving performance. Despite some pre-trained models being trained in the multilingual datasets, the English-translated pairs still perform better. For example, XLM-Roberta with and without

Model	Multilingual pre-trained	Pre-processing	Translated	F1	Recall	Precision
TwHIN-BERT	✓	✗	✗	0.3507	0.3585	0.3742
Bert	✗	✗	✗	0.2292	0.2461	0.3497
BERTweet	✗	✗	✗	0.2334	0.2493	0.3013
Roberta	✗	✗	✗	0.2334	0.2493	0.3013
XLM-Roberta	✓	✗	✗	0.2324	0.2466	0.2705
TwHIN-BERT	✓	✓	✗	0.2181	0.2319	0.3789
Bert	✗	✓	✗	0.2181	0.2319	0.3789
BERTweet	✗	✓	✗	0.2311	0.2450	0.3000
Roberta	✗	✓	✗	0.2311	0.2450	0.3000
XLM-Roberta	✓	✓	✗	0.2311	0.2450	0.3000
TwHIN-BERT	✓	✗	✓	**0.4581**	**0.4735**	**0.4460**
Bert	✗	✗	✓	0.4284	0.4345	0.4311
BERTweet	✗	✗	✓	0.3053	0.3123	0.3221
Roberta	✗	✗	✓	0.3053	0.3123	0.3221
XLM-Roberta	✓	✗	✓	0.3068	0.3079	0.3296
TwHIN-BERT	✓	✓	✓	0.4503	0.4696	0.4392
Bert	✗	✓	✓	0.4113	0.4175	0.4143
BERTweet	✗	✓	✓	0.3217	0.3331	0.3224
Roberta	✗	✓	✓	0.3217	0.3331	0.3224
XLM-Roberta	✓	✓	✓	0.3200	0.3268	0.3257

Table 6: The ablation study on pre-processing and translation. Translated indicated dev and test data translated into English via Google Translation API.

English translation, (0.3200/0.3268/0.3257) vs. (0.2311/0.2450/0.3000) for F1, Recall, and Precision, respectively. This is very intuitive because the training tweets are all in English. Therefore, models can only perform well on English tweet emotion classification.

A.5 Data-augmentation

Our results on data augmentation are shown in Table 7. As we can see, it achieved performance gain in almost every pre-trained model. Notably, this data augmentation only works well on BERT and TwHIN-BERT. It is harmful for BERTweet and Roberta pairs and almost zero-gain for XLM-Roberta.

A.6 In-Context Learning

A.6.1 Experiment Detail

We tested multiple OpenAI's GPT models using their APIs. Our system uses the latest model, GPT4o, which uses a different tokenizer that achieves better multilingual performance. Although ensuring deterministic output is challenging due to the GPU-based calculations of LLMs, we minimized the variables by setting the temperature, frequency penalty, and presence penalty as zero. Furthermore, we penalized undesirable output tokens and boosted the probability of desired labels. By setting log probability and max token number as one, we ensured our models to return label only.

A.6.2 Model Selection

We tested multiple GPT base models to find the best-performing one. On the dev dataset, Zeroshot GPT-4o outperformed both Zeroshot GPT-4 and Zeroshot GPT-3.5, achieving an F1 score of 0.5645, compared to 0.5616 and 0.4847, respectively. Therefore, we selected GPT-4o as our main model. Table 8 summarizes the results of various models and prompting methods.

A.6.3 Vulnerability of prompting

It is noteworthy that the effect of our Trigger Word CoT differs by model type. In Table 8, both GPT-4o and GPT-3.5 show a similar trend of increasing F1 scores with EngCoT and MulCoT. Compared to Zeroshot, EngCoT improved by approximately 0.2 to 0.4, and MulCoT gained roughly 0.5 to 0.7. However, as shown in Table 2, such gains were not transferable to the test data. Additionally, GPT-4 suffered from additional CoT steps, with the F1 score decreasing by approximately 0.5. This

Model	Multilingual pre-trained	F1	Recall	Precision
TwHIN-Bert	✓	**0.5010**	**0.5204**	**0.4943**
Bert	✗	0.4964	0.5057	0.4907
BERTweet	✗	0.3130	0.3236	0.3393
Roberta	✗	0.3130	0.3236	0.3393
XLM-Roberta	✓	0.3247	0.3259	0.3345

Table 7: The results on English back-translated.

	F1	Recall	Precision
Zeroshot GPT-4o	0.5645	0.5510	0.5881
EngCoT GPT-4o	0.5847	0.5885	0.5956
MulCoT GPT-4o	**0.6101**	**0.6106**	**0.6155**
Zeroshot GPT-4	0.5616	0.5713	0.5709
Fewshot GPT-4	0.5465	0.5509	0.5660
EngCoT GPT-4	0.5115	0.5130	0.5413
MulCoT GPT-4	0.5489	0.5498	0.5702
Zeroshot GPT-3.5	0.4963	0.5225	0.5053
FewShot GPT-3.5	0.4951	0.5176	0.5217
EngCoT GPT-3.5	0.5305	0.5433	0.5564
MulCoT GPT-3.5	0.5614	0.5822	0.5581

Table 8: The overall results of GPT results with varying models and prompting methods. Note that scores are not the 5-run average.

demonstrates that the effect of CoT and prompting methods depends on the model type, as well as the content and distribution of the dataset.

	F1	Recall	Precision
EngCoT GPT-3.5 (8)	0.5305	0.5433	0.5564
- 2 Neutral (6)	0.4434	0.4155	0.5870
+ 2 Emotion (10)	0.5062	0.5133	0.5494

Table 9: Results of GPT3.5 with varying number of CoT examples

Additionally, we conducted a simple ablation study on the number of CoT examples using GPT-3.5, which benefited the most from CoT with the dev data. Following the annotation guidelines, the current CoT steps consist of 8 examples: 5 emotion categories and 3 cases of the Neutral label. When we reduced the number of cases to 6 by removing two random Neutral examples, the F1 score dropped from 0.5305 to 0.4434. Adding two additional emotion-label examples decreased the F1 score to 0.5062.

This ablation study suggests that the number of CoT examples affects the results. However, as mentioned earlier, our findings are specific to the model types and dataset, making it difficult to generalize that 8 steps are the best. Furthermore, there is no strong evidence that our instructions significantly impacted the results, as several studies suggest that the quality and content of instructions do not matter much in ICL (Min et al., 2022; Webson and Pavlick, 2022; Wang et al., 2023).

A.6.4 Error Analysis

This section provides a label-wise analysis of the test data. Due to time and financial constraints, we were not able to perform a detailed ablation study and experiments on the test data.

F1	ZeroShot	Δ EngCoT	Δ MulCoT
Anger	0.7342	+0.0056	+0.0024
Fear	0.5125	**+0.0361**	**+0.0230**
Joy	0.5376	+0.0079	+0.0028
Love	0.4972	**+0.0107**	**+0.0215**
Neutral	0.6901	**+0.0194**	**+0.0160**
Sadness	0.5516	+0.0061	-0.0094
Macro Avg	0.5872	+0.0143	+0.0094

Table 10: Label-wise F1 scores for ZeroShot, EngCoT, and MulCoT models.

As mentioned in Section 4.2.2, the majority of changes in F1 occur in the Fear, Love, and Neutral labels. Table 11 and Table 12 show that CoT improves the F1 scores of these labels by increasing Recall at the expense of Precision. Given that CoT improves Recall, the performance gain on Fear, the most underrepresented label in the dataset, seems intuitive. Similarly, Love, the second most scarce label, potentially benefits from the robustness provided by CoT. The changes in the Neutral label suggest that the model might be following the reasoning steps detailed in our CoT examples, where we emphasize the lack of keywords in the Neutral examples. However, there is no clear evidence supporting this hypothesis. Therefore, further research

is needed to understand the effect of CoT reasoning steps.

Recall	ZeroShot	Δ EngCoT	Δ MulCoT
Anger	0.7378	-0.0407	-0.0228
Fear	0.5325	+0.0909	+0.1039
Joy	0.4873	-0.0092	+0.0069
Love	0.4684	+0.0421	+0.0790
Neutral	0.7380	+0.0393	-0.0153
Sadness	0.5148	+0.0222	+0.0445
Macro Avg	0.5798	+0.0241	+0.0327

Table 11: Label-wise recall for ZeroShot, EngCoT, and MulCoT models.

Precision	ZeroShot	Δ EngCoT	Δ MulCoT
Anger	0.7306	+0.0576	+0.0289
Fear	0.4940	-0.0042	-0.0317
Joy	0.5994	+0.0356	-0.0033
Love	0.5298	-0.0246	-0.0369
Neutral	0.6481	+0.0045	+0.0422
Sadness	0.5940	-0.0140	-0.0679
Macro Avg	0.5993	+0.0092	-0.0114

Table 12: Label-wise precision for ZeroShot, EngCoT, and MulCoT models.

A.6.5 Prompts

This section provides a list of prompts used in the ICL experiments and descriptions.

CoT examples are made up of 8 examples: 5 Emotion labels and 3 different cases of Neutral labels. For non-Neutral labels, we randomly selected tweets with token sizes between 5 and 10. We excluded lengthy tweets (> 10) as it increases the cost and context window of prompts. On the other hand, shorter tweets (< 5) do not provide enough context information. Therefore, we randomly selected from the given range of tweets. We also included tweets with emojis in the CoT examples as emojis also serve as trigger words in the given dataset. Similarly, we randomly selected from three different scenarios of Neutral labels: bot-like, opinionated, and fact-stating tweets.

Role: system
Content: Classify given tweets in the following 6 labels: Joy, Anger, Sadness, Love, Fear, and Neutral. Your answer should be a label.

Role: user
Content: @user Yea he found it hilarious afterwards

Table 13: Base Prompts Examples.

System: Classify given tweets in the following 6 labels: Joy, Anger, Sadness, Love, Fear, and Neutral. Your answer should be a label.

User: Tweet: @user Job well done ! ! ! 200
The trigger word is 'Job well', indicating positive emotion. Also, it is not toward a specific person, more like an enthusiastic and energetic reaction.
Assistant: Joy
User: Tweet: My hair is so flat I hate it
The trigger word is 'hate it', indicating negative and furious emotion.
Assistant: Anger
User: Tweet: Our house looks so sad without the Christmas lights
The trigger words are 'so sad' and a sad emoji at the end, indicating sorrow.
Assistant: Sadness
User: Tweet: Thank you to whoever wonder traded me a shiny dialga
The trigger words are 'thank you', indicating positive emotion. Also, it is toward a specific person, 'me'.
Assistant: Love
User: Tweet: Gotta Move Back Home #PanicIn4Words
The trigger word is 'Panic' in PanicIn4Words, indicating being scared.
Assistant: Fear
User: Tweet: # NowOnAir @user Ft . @user - Nobody Knows . Listen live on http
There is no clear trigger word. This is also a tweet from bots.
Assistant: Neutral
User: Tweet: @user I always do this and I don't care at the morning afterwards ..
There is no clear triggering word. Therefore, this is simply stating an opinion without any indication of emotion.'
Assistant: Neutral
User: Tweet: @user Hi Prashant , We dont have exact dates / timelines , but were working to roll it out to all eligible devices globally as quickly as possible . Stay tuned ! - Tim
There is no clear trigger word. This is simply stating a fact.
Assistant: Neutral

User: Give me all four of those hours . Cut nothing . http

Table 14: Chain of thought example with English instructions.

System: Classify given tweets in the following 6 labels: Joy, Anger, Sadness, Love, Fear, and Neutral. Your answer should be a label.

User: Tweet: @user Job well done ! ! ! 200
The trigger word is 'Job well', indicating positive emotion. Also, it is not toward a specific person, more like an enthusiastic and energetic reaction.
Assistant: Joy
User: Tweet: Meine Haare sind so platt, dass ich es hasse
The trigger word are 'es hasse', indicating a negative and furious emotion.
Assistant: Anger
User: Tweet: Notre maison a l'air si triste sans les lumières de Noël
The trigger words are 'si triste' and a sad emoji, indicating sorrow.
Assistant: Sadness
User: Tweet: Bedankt aan degene die zich afvraagt of hij mij een glimmende dialga heeft geruild
The trigger word is 'Bedankt', indicating a positive emotion directed towards a specific person.
Assistant: Love
User: Tweet: Tengo que regresar a casa #PánicoEn4Palabras
The trigger word is 'Pánico' in PanicIn4Words, indicating being scared.
Assistant: Fear
User: Tweet: # NowOnAir @user Ft . @user - Nobody Knows . Listen live on http
There is no clear trigger word. This is also a tweet from bots.
Assistant: Neutral
User: Tweet: @user I always do this and I don't care at the morning afterwards ..
There is no clear triggering word. Therefore, this is simply stating an opinion without any indication of emotion.'
Assistant: Neutral
User: Tweet: @user Hi Prashant , We dont have exact dates / timelines , but were working to roll it out to all eligible devices globally as quickly as possible . Stay tuned ! - Tim
There is no clear trigger word. This is simply stating a fact.
Assistant: Neutral

User: Give me all four of those hours . Cut nothing . http

Table 15: Chain of thought example with multilingual instructions.