WU_TLAXE at WASSA 2024 Explainability for Cross-Lingual Emotion in Tweets Shared Task 1: Emotion through Translation using TwHIN-BERT and GPT

Jon Davenport[1], Keren Ruditsky[1], Anna Batra[1], Yulha Lhawa[1], Gina-Anne Levow[1],
[1]Computational Linguistics, University of Washington,
{jmeld,krudit,batraa,yulha,levow}@uw.edu
Correspondence: jmdeld@uw.edu

Abstract

This paper describes our task 1 submission for the WASSA 2024 shared task on *Explainability for Cross-lingual Emotion in Tweets*. Our task is to predict the correct emotion label (Anger, Sadness, Fear, Joy, Love, and Neutral) for a dataset of English, Dutch, French, Spanish, and Russian tweets, while training exclusively on English emotion labeled data, to reveal what kind of emotion detection information is transferable cross-language (Maladry et al., 2024). To that end, we used an ensemble of models with a GPT-4 decider. Our ensemble consisted of a few-shot GPT-4 prompt system and a TwHIN-BERT system fine-tuned on the EXALT and additional English data. We ranked 8th place under the name WU_TLAXE with an F1 Macro score of 0.573 on the test set. We also experimented with an English-only TwHIN-BERT model by translating the other languages into English for inference, which proved to be worse than the other models.

1 Introduction

Cross-lingual emotion analysis is vital to identifying emotions across diverse languages and addressing challenges such as linguistic diversity and cultural differences. Our approach utilizes transfer learning and cross-lingual word embeddings (Xu et al., 2022) as introduced in models like TwHIN-BERT (Zhang et al., 2022) to handle these variations.

Research highlights the effectiveness of transformer-based models like GPT and BERT in capturing contextual nuances for accurate emotion recognition (Acheampong et al., 2021). Studies also show the potential of large language models like GPT and RoBERTa to enhance user interactions (Venkatakrishnan et al., 2023). Our experiments used an ensemble approach, featuring a GPT-4 decider, a few-shot GPT-4 prompt system, and a fine-tuned TwHIN-BERT system, trained on the EXALT data supplemented with additional English data to optimize cross-language emotion detection. We additionally experimented with an English-only TwHIN-BERT model.

2 System Description

Figure 1, our best performing system submitted to CodaLab, proved to be an ensemble model with GPT-4 as the decider. The decider generates a label based on each tweet's text, and two predicted labels, each provided by a different system. The systems that provided these competing predictions for the decider were on the one hand, a TwHIN-BERT (Zhang et al., 2022) fine-tuned model, and on the other hand, the results of a few-shot GPT-4 prompt system. The fine-tuned TwHIN model was trained on a processed dataset based on the provided training data and a supplemental English, emotion-labeled dataset from a past shared task, SemEval 2018 (Mohammad et al., 2018). Since our goal was to develop a multilingual system, we used NLLB machine translation (Costa-jussà et al., 2022) to generate parallel corpora for each of the test languages (Spanish, French, Dutch, Russian). Then we balanced our dataset by down-sampling classes with over 10,000 examples, and up-sampling classes with fewer than 10,000 examples.

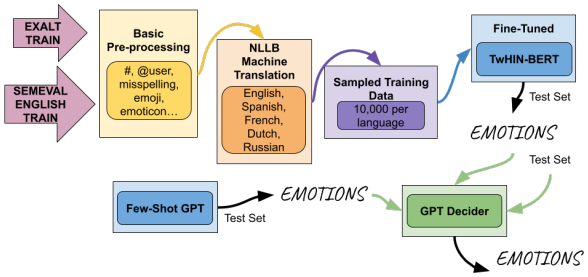

Figure 1: The GPT Ensemble architecture

We also experimented with an English-only

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 523–527

TwHIN-BERT model which uses NLLB to translate the other languages to English for inference, along with sampling of the data.

2.1 Few-Shot GPT-4

Few shot GPT-4 involved a simple system instruction prompt and a training set example for each language and each label. Chain of thought was not employed, as the goal was to evaluate GPT's performance in an information-rich prompt environment. We ran few shot GPT-4 on a computer cluster, using an API request. An example prompt is shown below.

```
{"role": "system", "content":
"You are a sentiment analysis
system assistant  designed to
classify the sentiment of each
tweet into one of the following
categories: Joy, Sadness, Love,
Anger, Neutral, Fear."},

{"role": "user", "content":
"J'ai téléchargé mon premier
article pour Gardez l'il
ouvert pour d'autres recettes
nutrition gainz delicious
biggestfan"},

{"role": "assistant",
"content":"Joy"}
```

2.2 Multilingual TwHIN-BERT

2.2.1 Dataset

The initial dataset consisted of the EXALT training data and relevant English samples from the SemEval 2018 task. We augmented that data by using NLLB to translate each English tweet to each target language (Spanish, Dutch, French, Russian). We then balanced the dataset such that each label had 10,000 samples, and split the balanced dataset into train (0.9) and development (0.1) subsets.

2.2.2 Approach

We fine-tuned the large version of TwHIN-BERT with our custom, multilingual dataset. We chose to use TwHIN-BERT because it has been pre-trained on tweets. We added a 6 label linear layer classification head to TwHIN-BERT, a dropout layer of 0.1, which then takes TwHIN-BERT's final layer's 1,024 dimension embedding and outputs a probability distribution over the label classes. For training, we set patience to two epochs and the learning rate at 2e-6, which resulted in 10 epochs of training, and used cross-entropy loss as our loss function. We fine-tuned TwHIN-BERT on a computer cluster with a L40 GPU.

2.3 GPT-4 Decider Ensemble

Our best performing multi-lingual model was a few-shot, ensemble model. We fed GPT-4 a system prompt that included instructions and several examples of an ensemble system decider. In each example the system is provided with the tweet's text, the label predicted by our multi-lingual fine-tuned TwHIN-BERT model, and the label predicted by our few-shot GPT query. An example prompt is shown below.

```
{"role": "system", "content":
"You are a sentiment analysis
ensemble classifier system
designed to classify the
sentiment of each tweet into
one of the following categories:
Joy, Sadness, Love, Anger,
Neutral, Fear. You will be given
a tweet and two labels provided
by other models, and you must
classify the sentiment based on
both the tweet and the other
model predictions."},

{"role": "user", "content":
"Label 1: Joy,  Label 2: Love,
Text: 15 year old tori-youve
been great. Bit of a twat but
youve been alright. Cant wait
to see the back end of you
tho ."},

{"role": "assistant",
"content":"Love"}
```

2.4 English-Only TwHIN-BERT

2.4.1 Dataset

We developed an English only dataset based on the EXALT training data and sampling valid tweets from the SemEval 2018 competition. This yields a much smaller dataset, which we then balanced in two different ways. We down-sampled to the least-represented emotion class ("Fear", count 616), and we also up-sampled to the most-represented emotion ("Joy", 2,933), and trained a model on each dataset. In order to run this model on the

training data, we developed a language-detection system, using spaCy, and then translated each (non-English) instance to English using NLLB. This resulted in an English-only version of the EXALT test set.

2.4.2 Approach

Throughout the competition, we were curious about how an English-only model would compare to a multi-lingual correlate. To test this, we experimented with an English only fine-tuned TwHIN-BERT. We fine-tuned the model in the same way as described with the Multilingual TwHIN-BERT (section 2.2).

3 Results

Table 1: Macro Test Results

Model	F1	Prec.	Recall
Few-Shot GPT-4	0.558	0.590	0.551
TwHIN-BERT ml	0.511	0.504	0.534
GPT-4 Decider	0.573	0.575	0.586
TwHIN-BERT en	0.440	0.447	0.495
Baseline	0.4476	-	-

The results in Table 1 show that the ensemble system with a GPT-4 decider achieved the highest performance with an F1 score of 0.573. Individually, the Few-Shot GPT-4 and TwHIN-BERT ml models scored lower, with F1 scores of 0.558 and 0.511 respectively. Thus, the ensemble method effectively enhanced the overall accuracy of emotion detection. These models also performed better than the EXALT organizer's baseline provided, which used inference on XLM-RoBERTa-base.

Our TwHIN-BERT en results, on the other hand, demonstrate that the English-only model performed worse than the rest, at 0.440. The TwHIN-BERT en, in fact, also performed slightly worse the organizer's baseline.

4 Discussion

As shown in the results section, the GPT-4 ensemble outperforms TwHIN-BERT and GPT-4 alone with respect to both F1 and recall. Few-shot GPT-4 had the highest precision and second highest recall and F1 scores followed by TwHIN-BERT. Focusing first on the three models attempted prior to the submission deadline, confusion matrices for all three models on the evaluation data are shown in Figures 2, 3, and 4. Compared to TWHIN-BERT, GPT-4 had higher accuracy in predicting Neutral (0.79 vs 0.57) and Fear (0.56 vs 0.39) labels, while TwHIN-BERT had better accuracy predicting Love (0.36 vs 0.49) and Sadness (0.45 vs 0.58) labels. Having access to decisions from both the previous models perhaps explains why the GPT-4 Decider model had the best performance with the highest accuracy across all labels aside from Sadness and Love (where TwHIN-BERT had the best results), and Neutral (where GPT-4 had the best results).

Improvement in the classification of Neutral labels appears to be the main contributor to the superior performance of the models making use of GPT-4 (GPT-4, GPT-4 Decider). However, the confusion matrices for the output of these two models also show relatively high rates of miss-classifying non-Neutral tweets as Neutral, suggesting that the GPT-4 models show a general over-reliance on the Neutral label. Given that there were more Neutral tweets in the evaluation data ('Neutral': 916, 'Joy': 433, 'Anger': 614, 'Sadness': 270, 'Fear': 77, 'Love': 190) compared to any other category, this also may account for the boost in performance seen by the GPT-4 models compared to TwHIN-BERT.

Across all three models, at least for the more common labels such as Joy and Anger, miss-classifications tended to cluster roughly by sentiment. For example, incorrect classifications of the Joy label were most often given to tweets labeled as Love or Neutral, and incorrect classifications of the Anger label were most often given to tweets labeled as Sadness or Fear. This suggests that even incorrect classifications often at least contained a similar sentiment (negative vs positive) to the actual label.

Moving on to the TwHIN-BERT model, the TwHIN-BERT en model performed far worse than all others on the test and development sets. The model was trained on an down-sampled dataset (TwHIN-BERT en), saw far fewer examples during fine-tuning than our multi-lingual model, and its performance suffered, proving to be our worst performing model.

A confusion matrix for the TwHIN-BERT en model is given below in Figure 5. TwHIN-BERT en manifests a bias towards predicting "Neutral" labels, despite training on a balanced dataset (each label had 616 samples). Consequently, it is possible that translation dilutes the intensity or polarity of some affect indicators. Perhaps more unexpected, is the same model's apparent tendency towards predicting "Sadness," and apparent aversion to predicting "Fear". We have fewer hypotheses for the

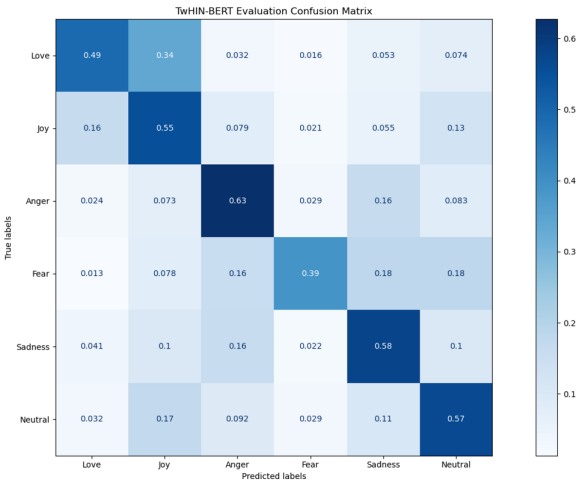

Figure 2: TwHIN-BERT ml Confusion Matrix

Figure 4: GPT-4 Decider Confusion Matrix

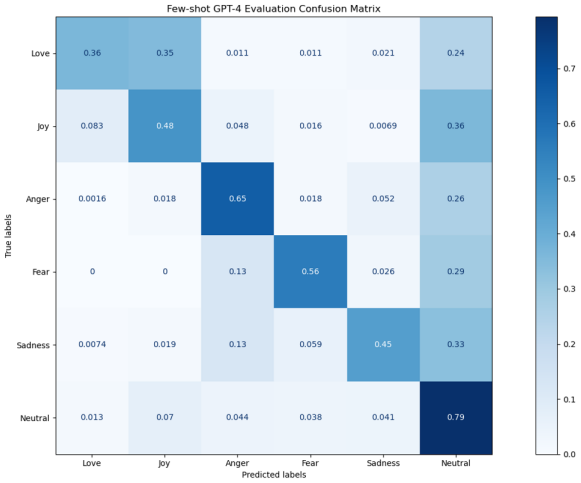

Figure 3: Few-shot GPT-4 Confusion Matrix

potential source of bias in these instances. Perhaps perceived "Fear" indicators are particularly challenging to translate, and perceived "Sadness" indicators are over-represented.

5 Conclusion

Novel, GPT-4 based models seem to out-perform straight-forward fine-tuning of the BERT based TwHIN-BERT even in few-shot contexts. The performance difference between the English-only and multi-lingual fine-tuned models surprised us. These results indicate that better future results might lie in prompt-based approaches to large language models. To that end, we foresee a wide range of experimentation in that domain, from chain of thought, to translation, multi-language prompting, and ensemble methods.

6 Limitations

We wanted to explore how fine-tuning a large language model like Llama-3 might perform, especially in comparison to few-shot GPT-4. Unfortunately, we could not acquire access to a GPU sufficient for that task in time. It seems possible, however, that ensemble and prompting techniques could prove more efficient or even superior to fine-tuning based approaches. We found late in our system building that continued pre-training on the test dataset domain, prior to fine-tuning, likely improves performance (Gururangan et al., 2020), and ideally we would like to test this approach as well.

Labeling emotions based off of short text, such as tweets, is highly subjective and it can be difficult to be consistent. This is a limitation of the shared task dataset and also extends to our model which is trained on this biased data.

One other notable limitation of our current systems is the reliance on translated datasets and data augmentation techniques that might not fully capture the nuanced expression of emotions across different languages and cultures. Translation errors and the inherent challenges of cross-lingual data can lead to misrepresentations of sentiment, affecting the models ability to accurately classify emotions in languages not originally included in the training set. This limitation highlights the need for better translation and data processing approaches that can more accurately reflect the true emotional content of different languages and cultural contexts.

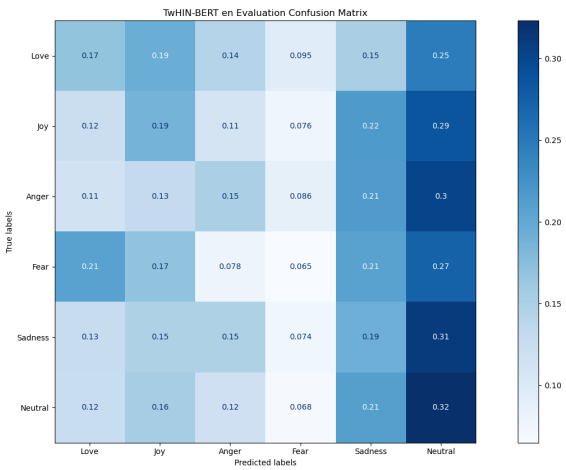

Figure 5: TwHIN-BERT en Confusion Matrix

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54:5789–5829.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Suchin Gururangan, Ana Marasovi, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

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*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

R. Venkatakrishnan, M. Goodarzi, and M. A. Canbaz. 2023. Exploring large language models' emotion detection abilities: Use cases from the middle east. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 241–244, Santa Clara, CA, USA. IEEE.

Yue Xu, Hua Cao, and Wei Du. 2022. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations. *Data Science and Engineering*, 7:279–299.

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. *arXiv preprint arXiv:2209.07562*.