PCICUNAM at WASSA 2024: Cross-lingual Emotion Detection Task with Hierarchical Classification and Weighted Loss Functions

Jesus Vázquez-Osorio [1,2], Gerardo Sierra[1,3],
Helena Gómez-Adorno[1,4], Gemma Bel-Enguix[1,3]

[1]Universidad Nacional Autónoma de México,
[2]Posgrado en Ciencia e Ingeniería de la Computación, [3]Instituto de Ingeniería,
[4]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Correspondence: jesusvo5599@comunidad.unam.mx, gsierram@iingen.unam.mx,
helena.gomez@iimas.unam.mx, gbele@iingen.unam.mx

Abstract

This paper addresses the shared task of multilingual emotion detection in tweets, presented at the Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA) co-located with the ACL 2024 conference. The task involves predicting emotions from six classes in tweets from five different languages using only English for model training. Our approach focuses on addressing class imbalance through data augmentation, hierarchical classification, and the application of focal loss and weighted cross-entropy loss functions. These methods enhance our transformer-based model's ability to transfer emotion detection capabilities across languages, resulting in improved performance despite the constraints of limited computational resources.

1 Introduction

This paper presents the team's proposal to solve the shared task 1 of multilingual classification of 6 emotions in tweets from 5 different languages using only English for model training. The presentation for this shared task was made for the Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA) that will be co-located with the Annual Meeting of the Association of Computational Linguistics (ACL) 2024 in Bangkok, Thailand (Maladry et al., 2024).

To address the task, the team focused on 3 methodologies for its resolution; the methodologies were mainly based on solving the imbalance of classes in the data. According to several works (Al-Azzawi et al., 2023), the increase of data, especially of the classes with fewer examples in the datasets, improves the result when performing the classification task with data not belonging to the training dataset, that is, the generalization of data in the models is improved. There are different methodologies for data augmentation in text classification tasks. As mentioned in (Shaikh et al., 2021; Edwards et al., 2023), and taking advantage of the latest advances in text generation, the use of generative language models is a great method for this data augmentation task.

In addition to data augmentation, a hierarchical ranking was also applied in the classification task in order to test the performance of the model with this methodology since, as shown in (Jr. and Freitas, 2011; Wang et al., 2022), this technique can result in great benefits in tasks with unbalanced data.

Finally, two loss functions, the focal loss and the cross-entropy weighted loss, introduced by (Lin et al., 2017) were also used, which allows focusing the training on difficult examples by reducing the contribution of well-classified examples, which is crucial to handling class imbalance.

2 Task Description

The task of emotion detection in tweets is a challenge in the field of Natural Language Processing (NLP) that explores the transfer of emotional information between languages. The sub-task 1 of the shared task of cross-lingual emotion detection task involves predicting emotions from six classes: Love, Joy, Anger, Fear, Sadness, and Neutral from tweets in five different languages Dutch, Russian, Spanish, English, and French.

A dataset of 5,000 pre-labeled English tweets is provided for training and 500 for validation, along with a test set of 2,500 tweets in the different target languages for evaluation. Participants may use additional English training resources to assess the effectiveness of the cross-language transfer approach but no other language different from English resources.

3 Methodology

First, during the training stage, only the 5,000 training tweets with their respective pre-labeled

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 490–494
August 15, 2024 ©2024 Association for Computational Linguistics

class were available, along with 500 validation data in different target languages without labels. In this stage, an exploratory analysis of the training data was carried out, in which the class imbalance in the training data was found. The target languages were detected in the validation data using the Python `langdetect` [1] library. In addition, using the `googletrans` [2] library, the texts were translated into English, and the emojis were converted into text with the `emoji` [3] library and the instances '*@user*' and '*http*' were removed from the tweets to make the predictions. Appendix B shows some examples of this text preprocessing.

3.1 Transformer Model Selection

Subsequently, using only the training data, different transformer models from the Hugging Face `Transformers` [4] library for zero-shot learning text classification were tested to predict the emotion of each tweet and obtain the classification report. Based on the results of these evaluations, the model with the highest macro *F1-score* in the training data test was selected to perform fine-tuning with the data.

3.2 Optimal Hyperparameter Search

Once the model to be used for this task was defined, the tokenizer of the pre-trained model was used to analyze the token length of the training tweets to define the token length to be used throughout the experiments. Next, a search for the best hyperparameters was conducted to fine-tune the model with the provided data. This hyperparameter search was performed using grid search, where the model was trained for one epoch with the training data split into 80% for training and 20% for validation, and different values for the hyperparameters '*weight decay*' and '*learning rate*' were proposed.

3.3 Strategies To Final Model

As mentioned, 3 techniques were used to handle class imbalance, which are described below:

3.3.1 Data Augmentation With Paraphrasing

With the training data, the '*humarin/chatgpt_paraphraser_on_T5_base*' model from Hugging Face, (Vladimir Vorobev, 2023), which was fine-tuned from the model T5-base from Google (Raffel et al., 2020) for text paraphrasing, was used. The data were augmented such that texts from the underrepresented classes, in this case, Love, Sadness and Fear, were duplicated to train the chosen classification model with these augmented data.

The training dataset is read, and texts labeled as Love, Sadness and Fear are extracted. Each of these texts is then tokenized and paraphrased using the pre-trained model. The paraphrased texts are added to a new dataframe along with their labels. This new dataframe is concatenated with the original dataset to create the augmented dataset. Appendix C shows some examples of the paraphrasing of texts using the above-mentioned model.

3.3.2 Hierarchical Ranking

Considering that there are 3 classes with the highest representation (Neutral, Joy, and Anger), and 3 with significantly lower representation in the training data (Love, Sadness, and Fear), the classification was trained with the chosen model in 2 stages. First, the model was trained to predict tweets among 4 classes: Neutral, Joy, Anger, and Other. Then, the same model was trained to predict among 3 classes: Love, Sadness, and Fear. The operation of this proposed technique involves performing the first classification (4 classes) and subsequently using the tweets classified as 'Other' as input for the second classifier (3 classes). This approach aims to prioritize the classification of the more represented classes.

3.3.3 Loss Functions

Considering the loss functions of (Lin et al., 2017), the chosen transformer model was trained by adapting these functions according to the class imbalance, as they assign different weights to the classes based on their representation. This increases the importance of the difficult-to-classify examples by adding smoothness to the class labels to demonstrate generalization with other data.

Once the training phase was completed, the labels for the validation data were released to continue evaluating models; additionally, the test data was released, which was also subjected to the translation process using the same methodology as the validation data to test the final models.

4 Results

Figure 1 shows the result of the analysis of the classes in the training set, it is evident that the rep-

[1] https://pypi.org/project/langdetect/

[2] https://pypi.org/project/googletrans/

[3] https://pypi.org/project/emoji/

[4] https://pypi.org/project/transformers/

resentation of the Love, Sadness and Fear classes is significantly lower.

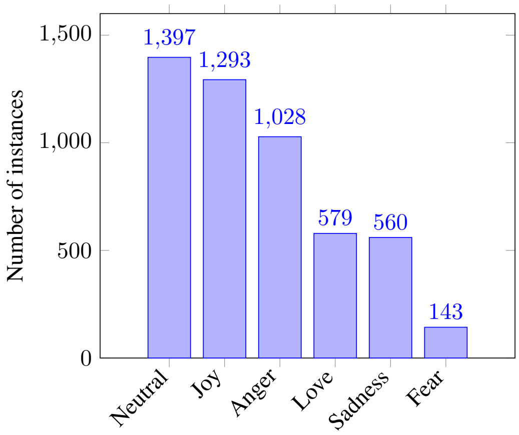

Figure 1: Class balance in training data provided.

Figure 2 presents the class distribution after data augmentation with model for text paraphrasing.

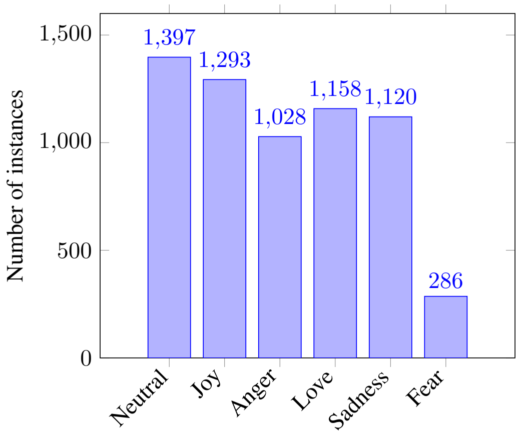

Figure 2: Class balance with data augmentation by paraphrasing data of the underrepresented classes.

For the selection of the transformer model, we present the highest results of the tested models in Table 1 (Sileod, 2022; AI, 2021). As mentioned, the model with the best performance evaluated with *F1-score* was selected, this is a fine-tuned model based on the *DeBERTa-v3-large* (He et al., 2021) from Microsoft, the model selected and used along all the experiments is '*MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli*' (Laurer et al., 2022) which was fine-tuned by us with the competition data.

In the grid search, the values from the Figure 3 were used for each hyperparameter in the table; within the search for the optimal pair of values for this task, *learning rate*=$5e-6$ and *weight decay*=0.01 emerged as the best options among the possible values.

Model	F1-score
sileod/deberta-v3-base-tasksource-nli	.39
facebook/bart-large-mnli	.40
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	**.45**

Table 1: Top three model performance without fine-tuning on training data.

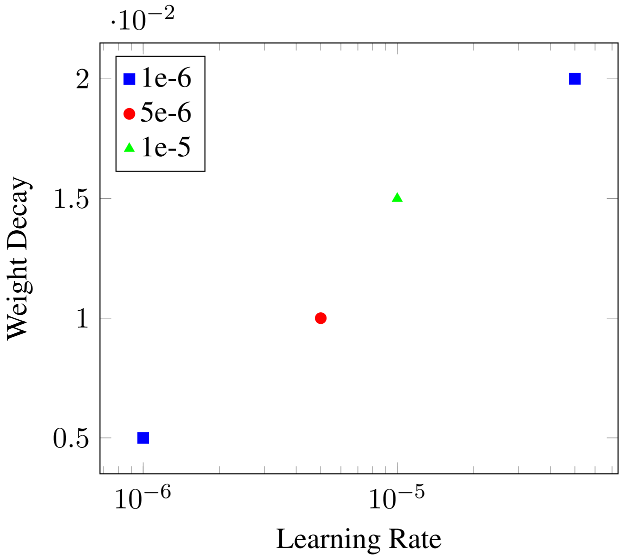

Figure 3: Values of the proposed hyperparameters to find the optimal pair.

For the fine-tuning of the model with the different strategies, different experiments were performed, not only 1 out of 1. Table 2 shows the best results obtained and the specifications of each one of them for the training phase.

For the evaluation phase we had the opportunity to present to the `CodaLab` [5] platform ten different predictions to get the performance of our models, in addition to testing the models in Table 2, based on the results of the training stage, for the evaluation stage the model with the loss function strategy was retrained with the same training data a few times to have different models trained and then used to rank the test data and present the predictions. Table 3 shows our top five prediction performances on the test data, which were obtained from the retraining of the model with the loss function strategy.

With the best result of 0.5183 we managed to beat the baseline provided by the organizers (0.4476). As for the participants, with this result

[5] `https://codalab.lisn.upsaclay.fr/competitions/17730`

Strategy	**Epochs**	***Original Val. Data****	***Preprocessed Val. Data*****
Without any strategy	10	0.4664	0.4669
Data augmentation	9	0.4716	0.4761
Hierarchical ranking	10	0.4935	0.4939
Loss functions	10	**0.5013**	**0.5063**

Table 2: Best results of the macro *F1-score* metric training stage for each strategy. *As provided. **Translated tweets, emojis converted to text and removing '*@user*' and '*http*' instances from tweets.

Epoch	***F1-score***
13	0.5022
17	0.5137
19	0.5168
15	**0.5183**
15	0.5099

Table 3: Best results of the macro *F1-score* metric evaluation stage.

we placed in the top 15.

5 Limitations and Future Work

The development of this task was carried out using limited computational resources (See A). For the training and evaluation of the models, the free resources of `Google Colab` [6] environment were used, supplemented with our own computational capacity. This restriction posed additional challenges, such as the need to optimize the use of available computing time and efficiently manage memory and processing resources. Despite these limitations, we were able to implement and experiment with advanced emotion classification models, demonstrating the feasibility of conducting significant NLP research with accessible and limited resources.

It is important to acknowledge that the hierarchical ranking approach may introduce cascading errors from the first classification stage to the second. This potential issue arises because any misclassification in the first stage (4 classes) can lead to incorrect input for the second stage (3 classes), thereby propagating errors. While this experiment did not include an in-depth study to evaluate the impact of these cascading errors, future work could focus on implementing and testing strategies to mitigate such issues. Possible solutions include using confidence thresholds to filter uncertain predictions, incorporating feedback loops for error correction, or employing ensemble methods to enhance the robustness of the hierarchical classification.

Unfortunately, due to time and resource constraints, we were unable to conduct an ablation study on the three techniques proposed in this paper. An ablation study would be valuable to isolate and compare the individual contributions of each technique to the overall performance. Future research should aim to conduct such a study to better understand the effectiveness of each technique when used separately and in conjunction with others. This would provide a clearer picture of the strengths and weaknesses of each approach and help optimize the overall classification performance.

6 Conclusion

In this work, we tackled the challenge of cross-lingual emotion detection in tweets using a transformer-based model trained only on English data. To overcome the class imbalance inherent in the dataset, we employed strategies such as data augmentation through paraphrasing, hierarchical classification, and the use of focal loss and weighted cross-entropy loss functions.

Despite utilizing limited computational resources, including free Google Colab environments and our own hardware, our approach demonstrated the feasibility of achieving competitive results in multilingual emotion detection tasks.

Environment	Google Colab
GPU	T4 GPU
GPU RAM	15 GB
System RAM	12.7 GB
CUDA Version	12.1
Transformers Library Version	4.40.2

Table 4: Software and hardware environment.

7 Acknowledgments

This work was carried out as part of PAPIIT project IT100822, IN104424 and the project CF2023-

[6] https://colab.research.google.com/

G64, which is funded by '*Consejo Nacional de Humanidades, Ciencias y Tecnologías* (CONAHCYT)'. We also extend our gratitude to our colleagues and mentors for their invaluable guidance and assistance throughout this project.

References

Facebook AI. 2021. Bart large for mnli.

Sana Sabah Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. Nlp-ltu at semeval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset. *arXiv*.

Aleksandra Edwards, Asahi Ushio, Hélène de Ribaupierre, Jose Camacho-Collados, and Alun Preece. 2023. Guiding generative language models for data augmentation in few-shot text classification. *arXiv preprint arXiv:2111.09064v2*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Carlos N. Silla Jr. and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating more clasifying - addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *Facebook AI Research (FAIR)*.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sarang Shaikh, Sher Muhammad Daudpota, Ali Shariq Imran, and Zenun Kastrati. 2021. Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models. *Applied Sciences*, 11(869).

Sileod. 2022. Deberta v3 base for tasksource nl.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *arXiv preprint arXiv:2203.03825*.

A Description of the computer resources used in the development of the task.

Table 4 shows the hardware and software environment with which all experiments were run.

B Examples of text translation.

Original Text: Wat een mega baas die @user Op HET moment het doen . Absurd goed dit . rtl7darts wkdarts

Translated Text: What a mega boss who @USER is doing it at the moment.Absurd well this. rtl7darts wkdarts

Original Text: bref je vais finir les 4 pages de mon livre on se retrouve quand je serais desséchée http

Translated Text: In short I will finish the 4 pages of my book we meet when I am dried up http

Original Text: Quien le mete papas fritas a los sándwiches de miga ? digo así somos amigos

Translated Text: Who puts french fries to crumb sandwiches?I say so we are friends.

C Examples of text paraphrasing.

Original Text: Mood of the day : worrying about online friends while being afraid of taking the risk of getting too close or too caring so not speaking while regretting to do so .

Paraphrased Text: The mood today is focused on stifling online friendships and the fear of losing too much control or attachment to others, leading to a lack of conversation and regret.

Original Text: @user But the fact your so hurt by the fact your idol has a boyfriend actually is homophobic much.

Paraphrased Text: The fact that your idol's partner is a homophobe is so hurtful to you.

Original Text: Gotta Move Back Home PanicIn4Words.

Paraphrased Text: I'm in a panic mode during the PanicIn4Words event, and it's time to move back home.