OZemi at SemEval-2025 Task 11: Multilingual Emotion Detection and Intensity

Hidetsune Takahashi **Sumiko Teng** **Jina Lee**

Wenxiao Hu **Rio Obe** **Chuen Shin Yong**

Emily Ohman
Waseda University
ohman@waseda.jp

Abstract

This paper presents the OZemi team's submission to SemEval-2025 Task 11: Multilingual Emotion Detection and Intensity. Our approach prioritized computational efficiency, leveraging lightweight models that achieved competitive results even for low-resource languages. We addressed data imbalance through data augmentation techniques such as back-translation and class balancing. Our system utilized multilingual BERT and machine translation to enhance performance across 35 languages. Despite ranking mid-tier overall, our results demonstrate that relatively simple models can yield adequate performance across diverse linguistic settings. We provide an error analysis of emotion classification challenges, particularly for nuanced expressions such as sarcasm and irony, and discuss the impact of emoji representation on model predictions. Finally, we outline future directions, including improvements in sentiment intensity modeling and the integration of semantic prosody to refine emotion detection.

1 Introduction

This paper explores SemEval 2025 task 11 (Muhammad et al., 2025b), which focuses on multi-label emotion detection and emotion intensity across various languages based on the datasets provided by the task organizers (Muhammad et al., 2025a; Belay et al., 2025). The task is divided into three tracks. Track A involves predicting the presence of emotion(s) such as joy, sadness, anger, surprise, and disgust in text snippets. Each emotion is labeled in a binary format: (1) if it is present, and (0) if absent. Task B focuses on predicting the intensity of a perceived emotion on a scale of 0 (*no emotion*) to 3 (*high intensity of emotion*). Finally, Track C is about using a trained dataset in one language to predict emotion labels in a different language. The datasets cover 35 languages in total, with genres ranging from social media to conversational text.

Emotions are at the core of human interactions but are notoriously difficult to detect in text (Öhman, 2021a). Any technical and theoretical advancements have the potential to aid in customer service automation, online content moderation, and many other tasks – both academic and commercial. The focus on cross-lingual emotion detection assists the recognition of emotions on a global scale, increasing its relevance among different cultural contexts.

Our team ranked around the middle for all tasks and languages. Our approach is not the most technically advanced, but because of this it is also not computationally very intensive. We managed to show that it is possible to achieve adequate results even for low-resource languages with very little computational resources. Our code is available on GitHub[1].

2 Background and Previous Work

The input for this task is text snippets in multiple languages ranging from commonly spoken languages such as English, German, and Spanish, to less commonly spoken languages such as Emakhuwa. The output differs across various tracks. The objectives of each Track are demonstrated using the sentence "I just won the lottery!" as an example.

Track A (Multi-label Emotion Detection): The output consists of binary labels for each perceived emotion, where (1) indicates its presence and (0) an absence. The output would look something like: Joy: 1, Sadness: 0, Fear: 0, Surprise: 1, Disgust: 0, Anger: 0 For some languages (such as English), the set of perceived emotions does not include Disgust.

Track B (Emotion Intensity): The output consists of an intensity prediction for each perceived emotion. The output would look something like: Joy: 3, Sadness: 0, Fear: 0, Surprise: 2, Disgust:

[1] https://github.com/esohman/SemEval2025/

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 109–115

0 The degrees of intensity range from 0 to 3, with 0 indicating no emotion, and 3 indicating a high intensity of emotion. The above output example for the sample sentence indicates a high intensity of Joy and moderate intensity for Surprise.

Track C(Cross-lingual Emotion Detection): Involves predicting emotion labels for a language using a labeled training set in a different language.

Track B also has additional challenges, not only does the emotion need to be accurately categorized, but labeled with intensity as well. As Kiritchenko and Mohammad (2017) state, rating scales used as annotations for sentiment analysis suffer from various flaws. They can be inconsistent, with gaps in value between annotators despite agreement on general sentiment, biased towards certain parts of the scale, or suffer from either too little or too much granularity. Additionally, further difficulty can emerge from methods of data pre-processing or lemmatization that may make data easier to categorize when using traditional methods. Exclamation marks, capitalization and more lend context to emotion intensity but may cause difficulty in processing. In addition, sarcasm and irony, prevalent in many common sources of data such as Twitter and other social media can also increase the difficulty of categorization and intensity mapping. However, emotion intensity is an important measure that has been shown to correspond well with human interpretations of a text's overall emotional content (Öhman, 2021b).

3 System Overview and Experimental Setup

In the research on the English model for Track A, we handled emojis by using the *demojize* function of the emoji Python library to convert emojis into descriptive textual labels (Kim and Wurster, 2025).

Track A and C use the f-score, and tack B Pearson correlation for evaluating the models.

For both Track A and Track B, the training data was split into two groups: training and testing sets. The preparation of the dataset involved data cleaning process to ensure the text inputs were uniform and to avoid unnecessary characters. This included replacing or removing special characters and standardizing representations for symbols such as quotes.

3.1 Data Imbalance

One of the significant challenges encountered during the experiment was the imbalance in the dataset's label distribution as shown in Table 1.

Emotion	Count
Anger	497
Fear	2,573
Joy	963
Sadness	1,376
Surprise	1,126

Table 1: Label distribution in the dataset.

The imbalance was most pronounced in the "Anger" class, which had substantially fewer samples than other categories. This posed a risk of bias during model training, as the model might underperform in recognizing emotions associated with underrepresented classes.

To address this issue, three strategies were employed:

1. **Data Duplication**

   Instances from the minority class ("Anger") were duplicated to match the sample size of the majority classes. This ensured that all emotion classes had equal representation in the dataset, reducing the risk of model bias.

2. **Synthetic Oversampling**

   For the Russian dataset, we explored more advanced sampling methods. Specifically, SMOTE (Synthetic Minority Oversampling Technique) was applied in combination with TF-IDF (Term Frequency-Inverse Document Frequency) to balance the class distribution. First, we transformed the text data into numerical form using TF-IDF vectorization. Using the numerical form, synthetic data can be created using SMOTE by interpolating between minority class data and their nearest neighbors. After applying SMOTE, the resampled data is in numerical form as well. To maintain consistency with the rest of the data, the TF-IDF features are transformed back to text for use. As we recognize that data duplication might lead to overfitting, where the model learns to recognize repeated patterns rather than generalizing well, SMOTE was used as an alternative approach.

3. **Back Translation**

   Back translation was applied as secondary approach. This technique involved translating the minority class samples into German or other languages and then translating them back into English using translation models such as DeepL and Google Translate. This method created syntactically diverse examples while preserving the semantic meaning of the original text, effectively augmenting the dataset with quality synthetic data.

These methods were implemented, and their effectiveness in balancing the dataset was evaluated during model training and testing.

3.2 Application and Model Enhancement

The following methodologies were applied in order to handle various languages and to enhance our fine-tuned model.

1. **External Data**

   External data[2][3] were used to examine whether or not they could have positive influences on results resolving the data imbalance discussed before. Judging from the texts, much of the data seems to have originated from social network platforms. The additional data were concatenated with the original dataset to balance the number of sentences between emotional/non-emotional for each of the emotions. By leveraging the balanced data, the BERT model was trained again in English and tested on the development dataset. The addition of the external datasets caused the results to drop from 66% to 55%. This indicates that the official data might be of higher quality annotation-wise or have some unique features compared to the external data.

2. **Hyperparameter Tuning**

   Hyperparameter tuning was implemented as our approach to enhance our model performance. The learning rate was adjusted from three values (2e-5, 3e-5, 5e-5), and the number of epochs was adjusted from 2, 3, and 4. We then chose the best performing parameter sets for each of the emotions based on development data and saved the weights trained with them.

3. **Machine Translation**

   Machine translation techniques were applied to implement our baseline in English and German for Track A and Track C. Google Translate was used as an example of our approach, leveraging its free availability in a multitude of languages. This technique was applied to all the task languages that Google Translate supports, which was 26 out of 28 and 30 out of 32 languages for Track A and Track C, respectively, at the time of writing. The languages not covered by Google Translate was Nigerian-Pidgin (PCM) and Emakhuwa (VMW), for which multilingual BERT was used to complement the limitation. This methodology was applied to leverage our German baseline for "disgust", and English baseline for the other emotions, since "disgust" is not included in the English dataset. This approach enables us to utilize our fairly strong baseline into various languages, additionally to analyze its impact on resource-wise both major and minor languages. Prior testing had also shown that language-specific models particularly for low-resource languages did not generally outperform multilingual ones (Takahashi et al., 2024) and thus the choice to stick to multilingual BERT for most languages was made.

4. **Language-Specific Models**

   However, besides the use of the machine translation technique discussed above, language-specific models were used for three non-English languages for better performances. We focused on Russian, German, and Chinese in this approach, leveraging the simple availability and proven robustness of language-specific BERT models. The models were fine-tuned by official training data and implemented into our system separately from the one that uses machine translation. This approach yielded mid-performing results in macro-F1 score in the official validation phase; 53 % for German and 54 % for Chinese. Comparing these results with those in other non-English languages, including 55 % of Afrikaans in the same phase, it can be said

[2] https://www.kaggle.com/datasets/parulpandey/emotion-dataset/data

[3] https://www.kaggle.com/datasets/nelgiriyewithana/emotions

	NRC Lexicon	Twitter Roberta Base
Anger	0.295	0.406
Fear	0.433	0.688
Joy	0.406	0.646
Sadness	0.457	0.612
Surprise	0.186	0.344

Table 2: Table with benchmark model F1 scores

that our trial to combine the fine-tuned model with machine translation has established a fairly good system for emotion detection over various languages, taking into account its simplicity and ease of deployment.

4 Results

We benchmarked our results against more simple models not fine-tuned on the specific instructions for this task. These benchmark models serve to provide points of reference to evaluate our more complex system. Our system proves its effectiveness and added value by achieving higher scores for this specific task.

One benchmark model uses the NRC Lexicon (Mohammad and Turney, 2013) as a simple rule-based approach that relies on a predefined set of words labeled with Plutchik's 8 core emotions (Plutchik, 1980), matching words to emotions without context or taking negations and valence-shifters into account. As this approach is the simplest, we expect our model that takes context into account to outperform it.

The other benchmark model was the CardiffNLP RoBERTa Base Sentiment multi-label model fine-tuned for SemEval 2018 task 1 (Camacho-Collados et al., 2022). This model is pre-trained on a large corpus of tweets and captures contextual word representations. However, it has not been fine-tuned for this specific task in this evaluation. Its performance already shows a significant improvement over the NRC Lexicon due to its ability to understand the semantics of language at a deeper level, showing us that our system can benefit from understanding the nuances specific to this task.

5 Limitations

One limitation we identified is the inaccuracies in the training data tags provided. For example, although the sentence "But not very happy" does not have the sentiment of joy, the training data had it labeled as as "joy =1." By fine-tuning a model to produce high accuracy scores with respect to an inaccurately tagged dataset, this model, or models produced for this task, may only be accurate for this task, but not when solving other real-world problems.

Additionally, the provided English dataset for Track A does not contain any emojis, which limited the opportunity to directly study the impact of emojis on emotion detection within the English language dataset.

Our simplistic approach to implement Google Translate as a machine translation technique might have had a slight influence on our inference. As stated by Takahashi et al. (2024), the original sentence and the sentence translated by Google Translate may differ in terms of semantic relations. Therefore, it is likely that some input sentences were semantically changed when its translation, and hence had a negative influence on the performance. Alternative ways including use of better-performing machine translation models were considered, but we chose the model on the basis of costs, assuming that the possible influence discussed above would not be significant compared to other factors.

6 Future Work

While brainstorming, we explored the application of semantic prosody (Sinclair, 1996) to the task. While not included in the final model, interesting trends were found that can be included in the refining of future models. Introduced by Sinclair (1996), a word's semantic prosody refers to its tendency to co-occur with specific sentiments or emotions. For example, the words "bring about" and "cause" have similar semantics, in terms of how they both speak about the reason behind an occurrence. Yet, "bring about" is more likely to co-occur with positive occurrences, while "cause" is more likely to co-occur with negative occurrences (McGee 2012; see Figures 1 and 2 for collocates of "bring about" and "cause" respectively, analyzed using AntConc (Anthony, 2024)). Therefore, "bring about"'s semantic prosody can be viewed as positive, and "cause"'s negative.

In line with this concept, we hypothesized that certain n-grams may occur more frequently with certain emotion tags than others. Hence, we ran an analysis for the most frequent 1-gram and 2-grams

[4]Corpus referenced is an American English corpus compiled by Potts and Baker (2012), with more than 1 million tokens.

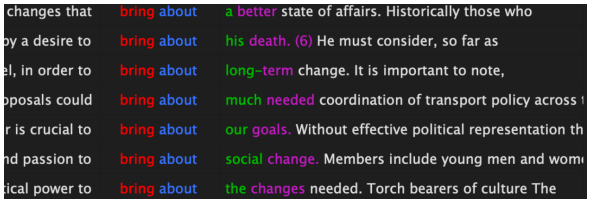

Figure 1: Key Word in Context (KWIC) of "Bring About"[4]

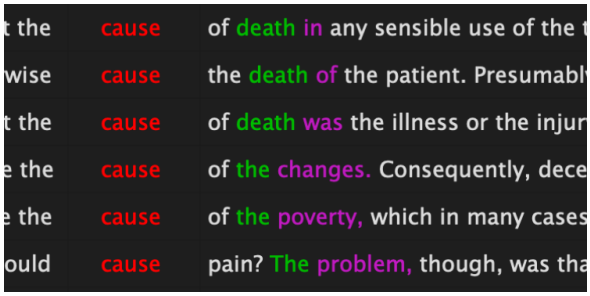

Figure 2: KWIC of "Cause"

for each emotion tag, stratified by each parts-of-speech tag. We found that there were unique n-grams for each emotion tag, occurring at a frequency of more than 1. It might thus be valuable to run this analysis with larger corpora, to find unique n-grams that co-occur with an emotion at a statistically significant frequency. These n-grams could then be used to further refine sentiment analysis models, especially multi-label ones that have relatively lower accuracy scores.

6.1 Emoji Analysis

To study the impact of emojis on emotion detection, we performed an emoji presence check on all language datasets in Track A. As shown in Table 3, German is the third most emoji-rich language in the dataset, with 255 occurrences. The first and second most emoji-rich languages are Somali and Sudanese, but compared to German, they lack sufficient BERT pre-trained models and other similar resources. Therefore, we chose German for our extension study. This additional study addresses a gap left by the English dataset, which does not contain any emojis.

We divided the German dataset into two groups: one containing emojis and the other with all emojis removed. For the emoji group, we used the demojize function to convert emojis into German text. We aimed to evaluate the impact of emojis by calculating the F-score for each model in these two groups.

As shown in Table 4, the F1 scores for the emoji

Language	Emoji Count
Somali (som)	373
Sundanese (sun)	363
German (deu)	255
Amharic (amh)	188
Tigrinya (tir)	33

Table 3: Top 5 Language in Track A by Emoji Count

	Emoji F1 Score	Non-Emoji F1 Score
Anger	0.952	0.955
Disgust	0.907	0.884
Fear	0.988	0.996
Joy	0.955	0.973
Sadness	0.966	0.977
Surprise	0.998	0.990

Table 4: F1 Scores for German Emoji and Non-Emoji Groups

and non-emoji groups are generally similar, with slight variations across different emotion labels. Apart from a slight improvement in the Disgust label with the inclusion of emojis, other labels, such as Fear, Joy, and Sadness, showed decreased performance when emojis were present. After conducting an error analysis on the emoji group, we found that the percentage of error texts with emojis was 15.79%, while that of error texts without emojis reached 84.21%.

Therefore, we conclude that although emojis can be beneficial for certain cases, such as improving the recognition of the *disgust* label, their overall impact on this German sentiment analysis model is limited and can sometimes negatively affect performance. However, it is also vital to note that analyzing the sentiment role of emojis is challenging due to reliance on context, cultural nuances, platform-specific features, and other related reasons (Hakami et al., 2022). Expanding this analysis to other languages could offer deeper insights into emojis' impact on sentiment and emotion detection.

7 Conclusions

Throughout prior discussions in this paper, it can be suggested that our straightforward approach under limited compute resources performs well even for low-resource languages. We have succeeded to maximize the benefits of lightweight models with experiments such as the use of back translation and hyperparameter tuning. Additionally, we have combined that with machine translation techniques

and also leveraged both multi-lingual and language-specific models over some non-English languages as well. In those ways, we firstly established our basis on emotion detection in English sentences, and then, applied the methodology to various languages. Although we had limited compute resources, this straightforward approach was shown to work well and be relatively competitive as well. Hence, our future work could also include some improvements in the same perspective, opening the door for its wider application to emotion detection in various languages.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K21058.

References

Laurence Anthony. 2024. Antconc (version 4.3.1) [computer software]. https://www.laurenceanthony.net/software/AntConc.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Shatha Ali A. Hakami, Robert Hendley, and Phillip Smith. 2022. Emoji sentiment roles for sentiment analysis: A case study in Arabic texts. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 346–355, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taehoon Kim and Kevin Wurster. 2025. Carpedm20/emoji: Emoji terminal output for python.

Svetlana Kiritchenko and Saif M Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv preprint arXiv:1712.01765*.

Iain McGee. 2012. Should we teach semantic prosody awareness? *RELC Journal*, 43(2):169–186.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhangand Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Emily Öhman. 2021a. *The Language of Emotions: Building and Applying Computational Methods for Emotion Detection for English and Beyond*. Ph.D. thesis, University of Helsinki.

Emily Öhman. 2021b. The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 7–12, NIT Silchar, India. NLP Association of India (NLPAI).

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Amanda Potts and Paul Baker. 2012. Does semantic tagging identify cultural change in british and american english? *International journal of corpus linguistics*, 17(3):295–324.

John Sinclair. 1996. The search for units of meaning. *Textus*, 9(1):75–106.

Hidetsune Takahashi, Xingru Lu, Sean Ishijima, Deokgyu Seo, Yongju Kim, Sehoon Park, Min Song, Kathylene Marante, Keitaro-Luke Iso, Hirotaka Tokura, et al. 2024. Ozemi at semeval-2024 task 1: A simplistic approach to textual relatedness evaluation using transformers and machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 7–12.

A Appendix

Performances in the test dataset are shown on Table 5. Four languages that are dealt with in Track C are not supported in Track A, to which the "*"s on the table correspond. The score for English on Track B is 0.6123, computed by the organizers' automatic calculation system just as in the other tracks.

language code	Track A	Track C
afr	0.4686	0.4708
amh	0.3701	0.3695
arq	0.4797	0.4793
ary	0.3395	0.3387
chn	0.4987	0.4987
deu	0.5082	0.5082
eng	0.6385	0.6385
esp	0.5251	0.5266
hau	0.3764	0.3778
hin	0.4686	0.4671
ibo	0.2850	0.2764
ind	*	0.4880
jav	*	0.4126
kin	0.2993	0.2989
mar	0.4979	0.4974
orm	0.3115	0.3118
pcm	0.4642	0.4642
ptbr	0.3456	0.3451
ptmz	0.2416	0.2442
ron	0.6422	0.6449
rus	0.7056	0.4398
som	0.3087	0.3098
sun	0.3994	0.4081
swa	0.2337	0.2320
swe	0.3829	0.3883
tat	0.4055	0.4037
tir	0.3306	0.3313
ukr	0.2791	0.2809
vmw	0.1931	0.1931
xho	*	0.3149
yor	0.2114	0.2109
zul	*	0.2121

Table 5: Official Performances on Test Dataset