NTA at SemEval-2025 Task 11: Enhanced Multilingual Textual Multi-label Emotion Detection via Integrated Augmentation Learning

Nguyen Pham Hoang Le[1,2], An Nguyen Tran Khuong[1,2], Tram Nguyen Thi Ngoc[1,2]
Thin Dang Van[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{22520982, 22520026, 22521498}@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

Emotion detection in text is crucial for various applications, but progress, especially in multi-label scenarios, is often hampered by data scarcity, particularly for low-resource languages like Emakhuwa and Tigrinya. This lack of data limits model performance and generalizability. To address this, the NTA team developed a system for SemEval-2025 Task 11, leveraging data augmentation techniques: swap, deletion, oversampling, emotion-focused synonym insertion and synonym replacement to enhance baseline models for multilingual textual multi-label emotion detection. Our proposed system achieved significantly higher macro F1-scores compared to the baseline across multiple languages, demonstrating a robust approach to tackling data scarcity. This resulted in a 17th place overall ranking on the private leaderboard, and remarkably, we achieved the highest score and became the winner in Tigrinya language, demonstrating the effectiveness of our approach in a low-resource setting.

1 Introduction

Emotions are influential in human interactions because it affects how people relate to each other, communicate, make decisions, and even sustain mental health. Being able to notice and understand emotions expressed in language is particularly important for the development of adaptive human-machine interfaces, AI systems with emotional intelligence, and targeted mental health care though combating depression. While there has been progress in emotion recognition, it is still deficient in most languages such as English, which happens to dominate the world. Leaving the bulk of the world's ethnolinguistic diversity unattended creates gaps that cannot be automatically filled by western emotional understanding. SemEval 2025 Task 11 “Bridging the Gap In Text-Based Emotion Detection"(Muhammad et al., 2025b) attempts to solve this problem using multilanguage emotion recognition for 32 African, Asian and European languages such as Emakhuwa, Amharic, Hausa and Swahili. The task mainly focuses on the recognition of insights emotions , that is, the emotion a listener infers from the speaker's words and takes into consideration the interactions surrounding it. The participants work in three tracks. Track A: Multi-label Emotions, track B: Emotion Intensity, track C: Cross-lingual Emotion Detection. We focus on Track A, Multi-label Emotions. We identify the emotion(s) portrayed by the speaker in the given target text snippet using Transformer models such as DeBERTa, RoBERTa, and mXLM-R, as these are powerful models from different linguistic families. To mitigate the lack of training data, we deploy augmentation strategies such as synonym replacement, deletion of random words, and oversampling for the minority emotion classes. These techniques improve model performance for languages with lower availability of resources. The analysis shows positive changes in macro-F1 scores, demonstrating the benefit of large language models in combination with advanced data for cross-cultural emotion detection. This not only solves the problem of language diversity, but also enhances the availability of emotion-sensitive technologies around the world.

2 Related Work

Emotion detection has evolved rapidly, fueled by transformer models (BERT, RoBERTa) that capture context better than old lexicon-based tools or handcrafted features(Mohammad and Turney, 2013). Multilingual efforts, like XLM-R(Conneau et al., 2020), struggle with low-resource languages—African languages, for instance, still lag despite adaptations like Afro-XLM-R(Alabi et al., 2022a). Adding to the chaos, multi-label emotion tasks require tweaked models to handle overlapping feelings. Data scarcity remains a hurdle.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1795–1801
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

While methods like EDA(Wei and Zou, 2019) randomly swap or delete words, our hybrid approach blends tailored augmentation with model adjustments, aiming to preserve cultural and linguistic quirks often lost in translation.

3 Shared Task Description

The SemEval-2025 Task 11 focuses on text-based emotion detection, specifically identifying the perceived emotion of a speaker based on a short text snippet. It consists of three tracks, however, we only focus on **Track A: Multi-label Emotion Detection:** Given a target text snippet, predict the perceived emotion(s) of the speaker. Specifically, select whether each of the following emotions apply: joy, sadness, fear, anger, surprise, or disgust. In other words, label the text snippet with: joy (1) or no joy (0), sadness (1) or no sadness (0), anger (1) or no anger (0), surprise (1) or no surprise (0), and disgust (1) or no disgust (0). Note that for some languages such as English, the set perceived emotions includes 5 emotions: joy, sadness, fear, anger, or surprise and does not include disgust. A training dataset with gold emotion labels will be provided for this track. The dataset for this shared task(Muhammad et al., 2025a; Belay et al., 2025) is drawn from five multilingual sources: social media posts, personal narratives, talks/speeches, literary texts, and news data, including both human-written and machine-generated content. This track comprises 28 languages from various countries in Africa, Asia, and Europe, including: Afrikaans (afr), Algerian Arabic (arq), Amharic (amh), Chinese (chn), Emakhuwa (vmw), English (eng), German (deu), Hausa (hau), Hindi (hin), Igbo (ibo), Kinyarwanda (kin), Marathi (mar), Moroccan Arabic (ary), Nigerian Pidgin (pcm), Oromo (orm), Brazilian Portuguese (ptbr), Mozambican Portuguese (ptmz), Romanian (ron), Russian (rus), Somali (som), Latin American Spanish (esp), Sundanese (sun), Swahili (swa), Swedish (swe), Tatar (tat), Tigrinya (tir), Ukrainian (ukr), and Yoruba (yor).

4 System Overview

The architecture of our system is illustrated in **Figure 1**. The pipeline is divided into five main steps: preprocessing, data augmentation, fine-tuning, voting scheme and threshold optimization. The detailed structure of the pipeline is depicted in the following subsections below.

4.1 Preprocessing

The process of preparing data before training improves model results by standardizing and cleaning input data. This research involves executing the following series of preprocessing steps:

1. **Lowercasing:** The text standardization process converts every character to lowercase across all languages to maintain consistency and simplify data complexity. For example: *The Brown Fox* becomes *the brown fox*. By converting all words to lowercase the model learns to recognize words based on their meaning instead of their case.

2. **Whitespace Removal:** We eliminate all unnecessary spaces to maintain uniform spacing throughout the text. The process removes leading and trailing whitespace from the text and merges multiple spaces between words into one space. The removal of excess whitespace guarantees consistent spacing throughout the text which prevents unwanted variations as demonstrated by the transformation of *" Hello world! "* to *"Hello world! "*.

3. **Lexical Normalization:** In English processing we apply lexical normalization as a supplementary preprocessing approach. The process converts non-standard word variations into their standard forms. The process of lexical normalization expands standard English contractions into their complete forms so *you'll* turns into *you will*, *they're* becomes *they are*, and *can't* changes to *cannot*. The text conversion process replaces slang terms and abbreviated codes with their full standard language equivalents like turning *ASAP* into *As Soon As Possible*.

4.2 Data Augmentation

In attempts to train powerful models on languages that have scarce resources, we face issues with general model data quality and paraphrasing context-embedded emotions. Our model incorporates and generalizes four fundamental techniques—swapping, deleting, over-sampling, and emotion specific synonym substitution. In addition, our model uses targeted synonym substitute per sentences for English. We will explain each technique and its motivation in detail as we go along.

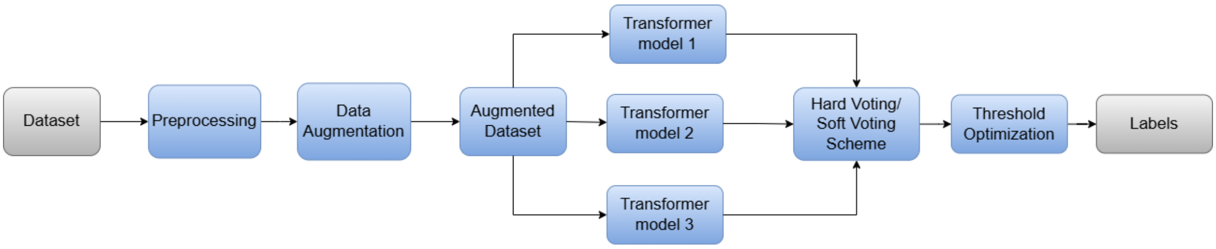

Figure 1: The architecture of the NTA team's system

4.2.1 Swap Augmentation

To address the problem of overreliance on fixed word order snares, we implement stochastic permutations, which consist of random positioning of two words in a sentence. This technique helps the model learn how to capture invariant emotional content regardless of the order of the words. For example, the sentence *"I love this beautiful place"* can be transformed to *"I beautiful this love place"*, thus, testing the ability of the model to discern emotional cues with respect to positional biases.

4.2.2 Deletion Augmentation

In most cases, real-world textual data is filled with noise or missing information. In order to recreate such conditions, we delete phrases within sentences that contain more than three words and this is where we let the model label emotions on its own. One example would be *"The weather is beautiful and sunny today"* being reduced to *"The is beautiful and today,"* and this checks capability of the model to classify accuracy when important words are absent.

4.2.3 Oversampling Augmentation

Oversampling augmentation is how we address class imbalance alongside motivating joint emotion recognition by creating new samples from existing ones. The label for the samples generated is set as the union of original sets of emotions present. For example, the combination of *"This news makes me surprised"* (surprise) and *"What a wonderful day!"* (joy) generates the sentence *"This news makes me surprised. What a wonderful day!"* with a composite label of *surprise* + *joy*. This technique enhances the model's ability to manage blended emotions simultaneously.

4.2.4 Emotion-Focused Synonym Insertion

To amplify emotional salience, we leverage the NRC Emotion Lexicon (Mohammad and Turney, 2013) to inject emotion-specific synonyms into sentences. For a sample labeled *anger* containing the word *"annoyed"*, we insert terms such as *"furious"* or *"enraged"* from the lexicon. This enriches emotional density without distorting the original sentiment, strengthening the model's grasp of domain-specific vocabulary.

4.2.5 Synonym Replacement

The synonym shifting process done in English is further aided by WordNet (Miller, 1994) allowing for more replacement while also being bound by emotion consistency checks via the NRC Lexicon. Consider, for example, *"I felt ecstatic after the celebration*." This text is transformed into *"I felt elated after the celebration*." Both the meaning and emotion intensity is retained. Such a technique can only be done in English because of the high availability of nuanced synonyms in WordNet that allow for generalization without losing accuracy in labels.

In conclusion, these methods expand the effective training corpus, promote robustness to syntactic variation, and sharpen the model's sensitivity to cross-lingual emotional signals—critical advantages for low-resource language processing.

4.3 Models

Our approach uses DeBERTa-v3, XLNet, and RoBERTa for processing English data. For non-English languages, we assemble an ensemble of multilingual models: mBERT, mXLM-R(for all non-English languages), and AfroXLMR-large (Alabi et al., 2022b) (for African languages). Model selection prioritized the ability to process diverse linguistic data and provide robust performance across a range of language tasks.

4.3.1 DeBERTa

DeBERTa(He et al., 2021) (Decoding-enhanced BERT with Disentangled Attention) is a model opportunity that enhances prorogation BERT and RoBERTa. In addition, it uses a disentangled attention mechanism whereby content and position addition are done separately. Thus, DeBERTa is superior because it has successfully proven to clarify the interaction between position and content of tokens in several NLP tasks.

4.3.2 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) refines BERT's architecture by training more extensively on a larger dataset and removing the next-sentence prediction mechanism, focusing solely on more robust language modeling tasks. This model has proven effective due to its optimized training approach and larger training corpus.

4.3.3 XLNet

XLNet(Yang et al., 2020) integrates Transformer-XL principles into a generalized autoregressive pre-training method. It outperforms BERT by learning on all possible permutations of the input sequence words, thus capturing a broader context and understanding the extended dependencies in text.

4.3.4 Multilingual Models

For tasks involving languages other than English, we harness the power of multilingual models:

- **mBERT**(Pires et al., 2019) (Multilingual BERT) is pretrained on a large corpus comprising text from 104 languages, which supports its capacity to understand and process multiple languages effectively.

- **mXLM-R** (XLM-RoBERTa) extends the capabilities of XLM models by training on an even more extensive multilingual dataset, further improving its effectiveness in cross-lingual settings.

- **AfroXLMR-large** was created by MLM adaptation of XLM-R-large model on 17 African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Nigerian-Pidgin, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu) covering the major African language families and 3 high-resource languages (Arabic, French, and English).

These models are pivotal in ensuring that our NLP solutions are not only efficient but also universally applicable across different linguistic backgrounds.

4.4 Ensemble Model

We experiment with two ensemble methods:

- **Soft-Voting:** Averages the prediction probabilities from each component model and selects the label with the highest average probability as the final output.

- **Hard-Voting:** Chooses the most frequently predicted label by the component models as the final output, applying a majority vote rule.

We use Hard-Voting for English and African languages and Soft-Voting for all other languages because Hard-Voting requires at least three models.

4.5 Threshold Optimization

The output logits of models are converted to probabilities using the sigmoid function. We adjust the threshold t for classification:

1. Iterate through potential threshold values from 0.40 to 0.60.

2. Compute the Binary Cross-Entropy loss for each threshold on the validation set.

3. Select the threshold t that minimizes the loss, optimizing the model's predictive accuracy.

This approach allows for fine-tuning the models to achieve optimal performance across diverse linguistic datasets and tasks.

5 Experimental Setup

- **Data and Pre-processing:** In our experiments, we utilize the training dataset provided by the organizers. We shuffle the data with a fixed random seed (42) and split it into training and validation subsets in an 80:20 ratio. This ensures a consistent and reproducible way to evaluate model performance throughout the development process.

- **Configuration Settings:** We implemented our models using the Trainer API from the Hugging Face library(Wolf et al., 2020) with the following hyperparameter settings:
  - **Learning rate**: 2e-5
  - **Optimizer**: AdamW with cosine learning rate scheduler
  - **Number of epochs**: 20
  - **Early stopping**: EarlyStoppingCallback with a patience of 3

6 Results

We present our results in two subsections. The first subsection compares results with and without augmentation, and the second compares official results and the SemEval Baseline.

6.1 Comparison of Augmented and Non-Augmented Results

The comparison of overall augmented and non-augmented results is shown in Table 1 (for the detailed results in each language, see appendix A). The augmented results achieve a better F1-Score in 21 out of 28 languages, and in the overall results, compared to the non-augmented results. These results prove that augmentation helps to increase the performance of the system, especially in low-resource languages.

Table 1: Comparison of Overall Augmented and Non-Augmented Results

Approach	**F1-Score**
Non-Augmented approach	0.4965
Augmented approach	0.5233
Comparison	+0.0267

6.2 Comparison of Official Results and SemEval Baseline

The official results are presented in Table 2. Our system achieves better results for the following languages: afr, amh, arq, chn, eng, hau, orm, som, sun, swe, tat, tir, vmw, and yor, compared to the SemEval baseline. Unfortunately, for the remaining languages, our system's performance was not as effective. We realized that this difference in performance is likely due to resource availability. For under-resourced languages, our augmentation learning helped the system improve performance, surpassing the baseline, remarkably, we achieved the highest score and became the winner in Tigrinya language, demonstrating the effectiveness of our approach in a low-resource setting. However, for rich-resourced languages, our augmentation learning did not significantly improve performance, and our model's performance was not as strong as the baseline's. Consequently, our results were lower than the baseline results.

Table 2: Our system's performance on private test sets

Language	**F1-Score**	**SemEval Baseline**
afr	**0.4073**	0.3714
amh	**0.6719**	0.6383
arq	**0.5063**	0.4141
ary	0.249	**0.4716**
chn	**0.5944**	0.5308
deu	0.5713	**0.6423**
eng	**0.761**	0.7083
esp	0.7528	**0.7744**
hau	**0.6783**	0.5955
hin	0.8367	**0.8551**
ibo	0.4703	**0.479**
kin	0.3798	**0.4629**
mar	0.8138	**0.822**
orm	**0.5048**	0.1263
pcm	0.4775	**0.555**
ptbr	0.3408	**0.4257**
ptmz	0.3459	**0.4591**
ron	0.6996	**0.7623**
rus	0.8347	**0.8377**
som	**0.4712**	0.4593
sun	**0.4198**	0.3731
swa	0.214	**0.2265**
swe	**0.5294**	0.5198
tat	**0.5694**	0.5394
tir	**0.5905**	0.4628
ukr	0.5335	**0.5345**
vmw	**0.2083**	0.1214
yor	**0.2188**	0.0922
Overall	**0.5233**	0.5093

7 Conclusion

In this study, we have taken on the dual challenge of multilingual emotion detection with multiple labels, specifically narrowing our focus on under-resourced languages. The hybrid approach that combined targeted data augmentation with language-specific model ensembles was found to be an effective approach, preventing further bias by the depletion of the data. The hybrid approach integrated lexical normalization, emotion-enriched synonym expansion, and adaptive threshold optimization, securing robust performance across various languages. Our system's performance beat the baseline in many languages and in overall. However, it is necessary to improve the performance in some languages. In future work, we plan to experiment more models and more augmentation techniques to increase the system's performance.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022a. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. *Preprint*, arXiv:2204.06487.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022b. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Preprint*, arXiv:1308.6297.

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 Zhang, 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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

A Detailed comparison of augmented and non-augmented results

In our preliminary study, we examine the performance of non-augmented and augmented approaches in each language. A detailed comparison of augmented and non-augmented results for each language and the overall results is presented in Table 3.

Table 3: Detailed comparison of augmented and non-augmented Results

Language	**Augmented F1-Score results**	**Non-augmented F1-Score results**
afr	**0.4073**	0.3336
amh	**0.6719**	0.6136
arq	**0.5063**	0.4727
ary	0.249	**0.3749**
chn	**0.5944**	0.5389
deu	**0.5713**	0.565
eng	0.761	**0.7626**
esp	**0.7528**	0.7434
hau	**0.6783**	0.5761
hin	0.8367	**0.8489**
ibo	0.4703	**0.4825**
kin	**0.3798**	0.3432
mar	0.8138	**0.8297**
orm	**0.5048**	0.4562
pcm	0.4775	**0.515**
ptbr	0.3408	**0.3619**
ptmz	**0.3459**	0.3346
ron	**0.6996**	0.6746
rus	**0.8347**	0.8318
som	**0.4712**	0.4341
sun	**0.4198**	0.3194
swa	**0.214**	0.211
swe	**0.5294**	0.4271
tat	**0.5694**	0.4936
tir	**0.5905**	0.442
ukr	**0.5335**	0.5267
vmw	**0.2083**	0.2034
yor	**0.2188**	0.1865
Overall	**0.5233**	0.4965