Pixel Phantoms at SemEval-2025 Task 11: Enhancing Multilingual Emotion Detection with a T5 and mT5-Based Approach

Jithu Morrison S **Janani Hariharakrishnan** **Harsh Pratap Singh**
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai - 603110, Tamil Nadu, India
{jithumorrison2210564, janani2210181, harshpratap2210854}@ssn.edu.in

Abstract

Emotion recognition in textual data is a crucial NLP task with applications in sentiment analysis and mental health monitoring. SemEval 2025 Task 11 introduces a multilingual dataset spanning 28 languages, including low-resource ones, to improve cross-lingual emotion detection. Our approach utilizes T5 for English and mT5 for other languages, fine-tuning them for multi-label classification and emotion intensity estimation. Our findings demonstrate the effectiveness of transformer-based models in capturing nuanced emotional expressions across diverse languages.

1 Introduction

Emotion recognition in textual data is a crucial task in natural language processing (NLP), with applications in sentiment analysis, mental health monitoring, and human-computer interaction. However, detecting emotions across multiple languages presents significant challenges due to linguistic diversity, cultural differences, and limited resources for many languages. SemEval-2025 Task 11, Bridging the Gap in Text-Based Emotion Detection, aims to improve emotion detection by providing a multilingual dataset covering 28 languages, including low-resource ones.

Our approach focuses on Track A (Multi-label Emotion Detection) and Track B (Emotion Intensity Estimation). For Track A, we fine-tune mT5 to classify multiple perceived emotions (joy, sadness, fear, anger, surprise, and disgust) within a given text. The model processes multilingual input and predicts relevant emotions simultaneously. For Track B, we extend this model to predict emotion intensities on an ordinal scale, ensuring that the system can accurately gauge varying degrees of emotional expression. This helps capture subtle differences in emotional intensity across languages.

During the task, we observed that multilingual emotion detection remains challenging due to variations in emotional expression and imbalanced datasets for low-resource languages. Our results show that transformer-based models like T5 and mT5 effectively capture emotional nuances, but performance varies depending on data availability. A key struggle was handling subjective emotional interpretations and linguistic inconsistencies across languages. Despite these challenges, our findings highlight the potential of multilingual models in improving cross-lingual emotion recognition.

2 Related Works

Emotion detection in text has been a long-standing challenge in NLP, evolving from traditional machine learning methods to deep learning and transformer-based models. Early approaches relied on feature engineering with statistical models such as SVMs and Naive Bayes, leveraging sentiment lexicons and syntactic dependencies (Cambria, 2017). The rise of deep learning introduced LSTMs, GRUs, and CNNs, which enhanced contextual understanding (Peters et al., 2018).

However, transformer-based architectures like BERT, RoBERTa, and T5 revolutionized the field with self-attention mechanisms and large-scale multilingual pretraining, significantly improving multi-label emotion classification (Devlin et al., 2019; Raffel et al., 2020). The growing interest in multilingual emotion detection has led to datasets like BRIGHTER (Muhammad et al., 2025a) (Muhammad et al., 2025b), which provide high-quality human-annotated emotion recognition data across 28 languages.

Our work builds on these advancements by applying T5 and mT5 models to SemEval-2025 Task 11, addressing multilingual emotion detection

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 617–622
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

challenges in two tracks: multi-label classification (Track A) and emotion intensity estimation (Track B). Unlike previous studies that focused primarily on high-resource languages, we fine-tune models on diverse linguistic datasets, leveraging BRIGHTER and other multilingual resources.

Additionally, our approach refines multi-label classification by using separate label schemas for different languages and tasks, setting it apart from standard transformer-based emotion classification models (Belay et al., 2025). Our experimental setup ensures robust evaluation across different linguistic contexts, further enhancing emotion understanding in low-resource languages (Baziotis et al., 2018; Mohammad et al., 2018). These advancements contribute to the broader goal of improving cross-linguistic emotion recognition and fostering more inclusive AI-driven applications in natural language processing.

3 System Overview

3.1 Key Algorithms and Modeling Decisions

The system is built on transformer-based architectures, primarily T5 and mT5, for multilingual text-based emotion detection. These models were selected due to their encoder-decoder design, which enables effective handling of both classification and regression tasks within a unified framework. T5 is well-suited for English text processing, while mT5, pretrained on a multilingual corpus, is optimized for handling diverse languages, including low-resource ones. This makes mT5 a strong candidate for cross-lingual emotion detection tasks. The model is fine-tuned for multi-label classification using a sigmoid activation function, allowing independent emotion predictions. Binary Cross-Entropy (BCE) is used for classification, while an ordinal regression loss captures the ordered nature of emotion intensities. Training is optimized using the AdamW optimizer with weight decay, and early stopping is applied to prevent overfitting.

3.2 Resources Beyond Training Data

Beyond the labeled training data, additional resources were incorporated to improve model performance. SentencePiece was used for tokenization, ensuring compatibility with multilingual input. Pretrained embeddings from the Hugging Face Transformers library provided a robust initialization for transfer learning. Weighted loss functions were used to address class imbalance, and external lexicons for sentiment analysis were explored to enhance the contextual understanding of emotions.

3.3 Mathematics Behind the Model

The transformer-based model follows a sequence-to-sequence architecture with self-attention mechanisms, allowing it to capture long-range dependencies and contextual cues across languages. For multi-label classification, the model computes the probability $P(y \mid x)$ for each emotion label independently as follows:

$$P(y \mid x) = \sigma(W \cdot h + b)$$

Here, $h \in R^d$ denotes the hidden state representation output by the final layer of the transformer for the [CLS]-like token (or averaged representation of the input), $W \in R^{k \times d}$ and $b \in R^k$ are learnable weights and biases, and σ is the sigmoid activation function applied element-wise to produce a probability distribution over the k emotion labels. Binary Cross-Entropy (BCE) loss is used for training, treating each label as an independent binary classification task.

For emotion intensity estimation, the model employs an ordinal regression framework to account for the ordered nature of intensity levels. Labels are encoded on a scale from 0 to 3, corresponding to increasing degrees of emotional intensity. Instead of treating intensity as a simple regression or multi-class classification problem, a cumulative link model is used, where the model learns a set of ordered thresholds $\theta_1 < \theta_2 < \ldots < \theta_{C-1}$ that separate adjacent ordinal classes. The probability that an input x belongs to class c is modeled as:

$$P(y \leq c \mid x) = \sigma(\theta_c - f(x))$$

where $f(x)$ is the scalar output from the model representing the underlying emotion intensity, and σ is the sigmoid function. This formulation preserves the ordinal structure of labels and ensures monotonicity across intensity levels. The corresponding loss is computed using the cumulative probabilities over all ordinal thresholds, optimizing the model to predict the correct ordered class.

This approach enables the model to better capture subtle variations in emotional intensity, particularly important in multilingual contexts where

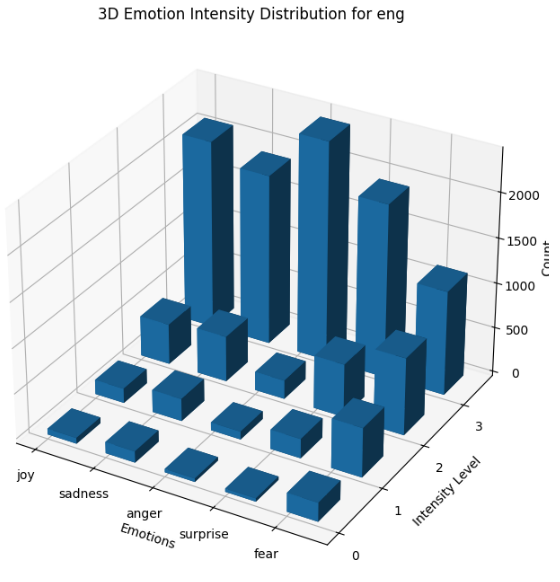

Figure 1: 3D Emotion Intensity Distribution for eng.csv

cultural nuances influence how emotions are expressed in text.

3.4 Variants of the Model

- **Track A (Multi-label Emotion Detection):**
  - **T5 Model:** Used for English text classification.
  - **mT5 Model:** Fine-tuned separately for two settings: five-label classification for African languages and six-label classification for other languages.
- **Track B (Emotion Intensity Estimation):**
  - **T5 Model:** Used for English text classification.
  - **mT5 Model:** Used for ordinal regression with six labels across all languages.

4 Experimental Setup

4.1 Data Splits and Usage

The dataset provided by SemEval-2025 Task 11 included a predefined test set. The training data was split into training and validation subsets to optimize hyperparameters and prevent overfitting. The test set remained untouched throughout model training and was only used for the final evaluation.

4.2 Preprocessing and Parameter Tuning

Text data underwent normalization steps such as lowercasing, whitespace trimming, and removal of special characters. Tokenization was performed using SentencePiece, ensuring effective encoding of multilingual text. Weighted loss functions were applied to mitigate class imbalances. Hyperparameter tuning was conducted using grid search, focusing on learning rate, batch size, and weight decay. Training was performed for 20 epochs with early stopping based on validation loss.

4.3 Model Architecture and Training Parameters

The models for both Track A and Track B were fine-tuned using T5 for English and mT5 for multilingual data. For Track A (Multi-label Emotion Detection), the T5 model was fine-tuned with a classification head using a sigmoid activation function for multi-label prediction. The mT5 model was used for multilingual settings, with a five-label classification schema for African languages and a six-label schema for all other languages. Both models utilized an AdamW optimizer with a learning rate of 5×10^{-5}, batch size of 8, and weight decay of 0.01. Early stopping was applied based on validation loss to prevent overfitting. For Track B (Emotion Intensity Estimation), mT5 was exclusively used with an ordinal regression framework to preserve intensity relationships. The same optimizer settings were applied, but the loss function was adjusted to accommodate ordinal regression. All models were trained for 20 epochs with checkpoints saved at each validation step to ensure robustness.

4.4 External Tools and Libraries

The implementation utilized several external tools and libraries to enhance efficiency and performance. PyTorch was used as the deep learning framework for model training and optimization. Tokenization and pretrained models were sourced from the Hugging Face Transformers library. Data processing was handled using pandas and NumPy, ensuring efficient manipulation of text and label distributions. For evaluation, Scikit-learn was employed to compute precision, recall, F1-score, and Pearson correlation. Training progress was monitored using tqdm, while hyperparameter tuning was conducted with Optuna to optimize learning rate and regularization parameters.

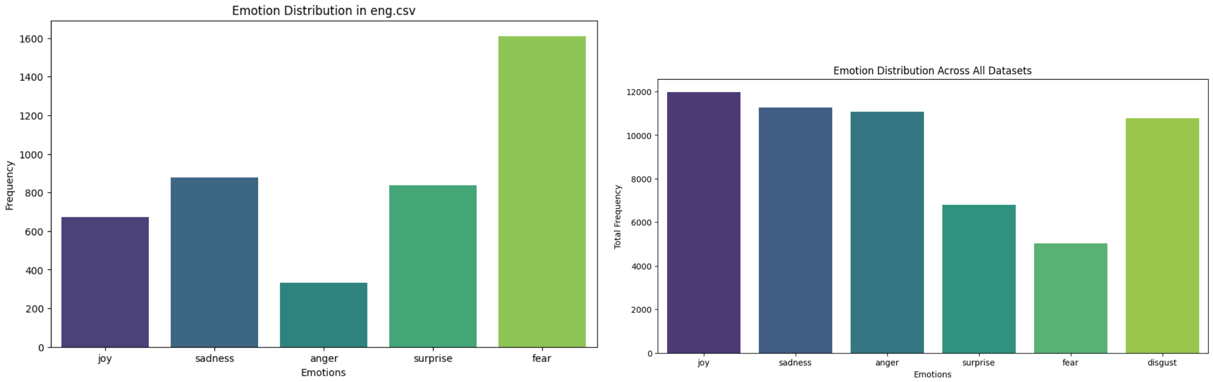

Figure 2: Track A - Emotion Distribution in eng.csv and all datasets respectively

5 Results

5.1 Track A: Multi-label Emotion Detection

The results for Track A highlight the effectiveness of the proposed approach across 28 languages, with model performance largely dependent on the availability and quality of training data. High-resource languages such as Hindi (hin), Marathi (mar), and Russian (rus) achieved strong F1-scores of 0.7304, 0.702, and 0.7205, respectively. This demonstrates that mT5, when fine-tuned on well-annotated datasets, can successfully classify multiple emotions in a multilingual setting. English (eng) also performed well with an F1-score of 0.5969, suggesting its stable position in multilingual pretraining.

However, performance drops significantly for low-resource languages such as Swahili (swa), Yoruba (yor), and Makhuwa (vmw), with F1-scores of 0.1775, 0.1473, and 0.0784, respectively. These results emphasize the challenges in generalizing to languages with limited annotated data. In such settings, the model struggles with sparse linguistic representation during pretraining and insufficient examples of emotion-labeled instances, which hinders its ability to learn meaningful associations. Additionally, variation in emotional expression across cultures and lack of task-specific linguistic resources further impact performance in these languages.

Interestingly, Spanish (esp) achieved an F1-score of 0.6403, despite being a non-high-resource language in the task. This indicates that factors such as annotation consistency, data diversity, and structural linguistic properties can significantly influence performance. Other moderately performing languages include Hausa (hau, 0.5015), Amharic (amh, 0.4371), and Romanian (ron, 0.5763), demonstrating that mT5 can still yield usable predictions in low-to-mid-resource settings if enough training signals are available.

Language	F1	Language	F1
afr	0.3063	pcm	0.4093
amh	0.4371	ptbr	0.2746
arq	0.3212	ptmz	0.2083
ary	0.3422	ron	0.5763
chn	0.3959	rus	0.7205
deu	0.3786	som	0.2541
eng	0.5969	sun	0.3095
esp	0.6403	swa	0.1775
hau	0.5015	swe	0.3893
hin	0.7304	tat	0.406
ibo	0.4102	tir	0.3198
kin	0.3167	ukr	0.353
mar	0.702	vmw	0.0784
orm	0.319	yor	0.1473

Table 1: Track A F1-score metrics

Languages such as Afrikaans (afr, 0.3063), Oromo (orm, 0.319), and Kinyrwanda (kin, 0.3167) struggled to surpass 0.35 F1, reiterating the difficulty of modeling underrepresented languages with minimal data. German (deu, 0.3786) and Swedish (swe, 0.3893), though not traditionally low-resource, underperformed, possibly due to limited or noisy annotations in the provided dataset.

5.2 Track B: Emotion Intensity

For Track B, where emotion intensity estimation was evaluated using Pearson correlation, the results similarly reflect a divide between high- and

Language	F1
amh	0.3769
arq	0.1119
chn	0.3656
deu	0.3424
eng	0.3285
esp	0.5279
hau	0.467
ptbr	0.2497
ron	0.406
rus	0.7448
ukr	0.2482

Table 2: Track B Pearson correlation

low-resource languages. Russian (rus) showed the highest correlation (0.7448), followed by Spanish (esp) with 0.5279. These results suggest that the model was able to rank emotional intensity reasonably well when sufficient training data was available.

However, substantial drops were observed for Algerian Arabic (arq, 0.1119), Ukrainian (ukr, 0.2482), and Brazilian Portuguese (ptbr, 0.2497), indicating challenges in estimating emotion intensity accurately under low-resource and linguistically diverse conditions. English (eng) exhibited a relatively low correlation (0.3285), likely due to the subtlety and ambiguity of emotional cues in English, which require deeper context-aware modeling. On the other hand, Hausa (hau) achieved a moderate score of 0.467, suggesting that with minimal but high-quality annotations, even low-resource languages can benefit from transformer-based fine-tuning.

Overall, the findings from both tracks demonstrate the potential of transformer-based models for multilingual emotion detection. However, they also expose clear limitations when applied to languages with limited or noisy datasets. Future work should focus on improving the representation of low-resource languages through transfer learning techniques, culturally aware embeddings, and enriched training datasets. Furthermore, integrating external resources such as emotion lexicons, morphological analyzers, and idiomatic expression banks may help bridge the gap in generalization across culturally and linguistically diverse settings.

6 Conclusion

This study demonstrated the effectiveness of transformer-based models, particularly T5 and mT5, for multilingual emotion detection. These models leveraged pre-trained knowledge and

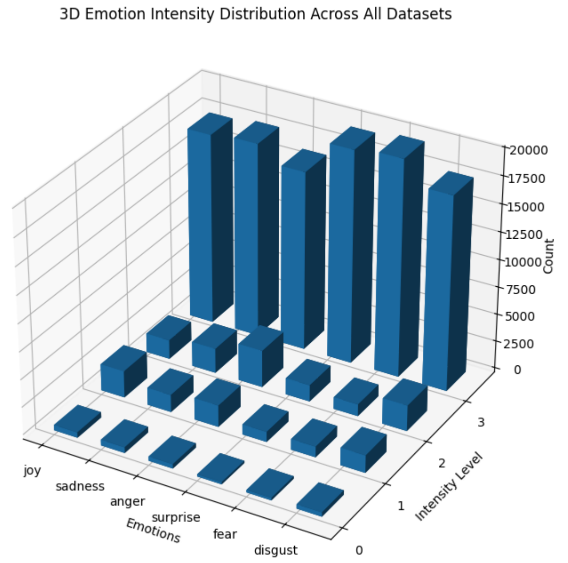

Figure 3: 3D Emotion Intensity Distribution across all datasets

fine-tuning to classify emotions across diverse languages, achieving strong results in high-resource languages like Hindi, Marathi, and Russian while facing challenges in low-resource languages such as Swahili, Yoruba, and Makhuwa due to data scarcity. The strong performance of Spanish, despite not being a high-resource language, suggests that factors like annotation quality and linguistic structure significantly impact model effectiveness. Emotion intensity estimation followed similar trends, highlighting the necessity of refined annotations and better training data. Future work should focus on enhancing dataset availability, optimizing model fine-tuning, and incorporating external linguistic resources to improve cross-linguistic performance and generalizability.

While the study focused on T5 and mT5 due to their unified text-to-text architecture, future work should also explore competitive encoder-only models like XLM-R. Additionally, leveraging transfer learning techniques, such as adapter layers or language-specific fine-tuning, could further improve low-resource performance.

7 Ethical Considerations

The development of multilingual emotion detection models presents ethical challenges, including biases due to uneven language representation, potential misinterpretations across cultures, and privacy concerns related to user-generated content. The performance gap between high- and low-resource languages risks marginalizing underrepresented communities, while cultural variations in emotional expression may lead to inaccurate predictions. Ensuring transparency, improving dataset diversity, and implementing robust privacy safeguards are essential to mitigate these risks. Collaboration with linguists, ethicists, and cultural experts can further refine these models to be more inclusive and ethically responsible.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2018. Ntua-slp at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 245–251. Association for 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.

Erik Cambria. 2017. Affective computing and sentiment analysis. In *IEEE Intelligent Systems*, volume 32, pages 102–107. IEEE.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Saif M 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. Association for Computational Linguistics.

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-2025 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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237. Association for Computational Linguistics.

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.