Stanford MLab at SemEval-2025 Task 11: Track B–Emotion Intensity Detection

Joseph Le and Hannah Cui and James Zhang
Stanford University

Abstract

We here outline our SemEval 2025 Track B: Emotion Intensity Prediction submission, for which the objective is to predict the intensity of six primary emotions—anger, disgust, fear, joy, sadness, and surprise—between 0 and 3, with 0 being none and 3 being very strong. We used a regression fine-tuned BERT-based model that makes use of pretrained embeddings in order to sense subtle emotional wordings in text.

We include tokenization with a BERT tokenizer, training with AdamW optimization, and an ExponentialLR scheduler used for learning rate modification. Performance is monitored based on validation loss and accuracy through closeness of model outputs to gold labels.

Our best-performing model is 68.97% accurate in validation and has a validation loss of 0.373, demonstrating BERT's capability in fine-grained emotion intensity prediction. Key findings include that fine-tuning transformer models with regression loss improves prediction accuracy and that early stopping and learning rate scheduling avoid overfitting. Future improvements can include larger datasets, ensemble models, or other architectures such as RoBERTa and T5. This paper shows the potential of pretrained transformers for emotion intensity estimation and lays the groundwork for future computational emotion analysis research.

1 Introduction

The SemEval 2025 Task 11 Track B: Emotion Intensity Prediction seeks to create models that make predictions about the perceived intensity of six emotions—joy, sadness, fear, anger, surprise, and disgust—in a sentence. The intensity is on an ordinal scale of 0 (no emotion) to 3 (strong emotion), which allows us to have a more nuanced view of emotional expression. This task is essential to the creation of emotion-aware NLP applications, such as sentiment analysis, mental health tracking, and human-computer interaction, by identifying not just the presence of emotion, but also its intensities. The dataset comprises eleven languages—Amharic, Algerian Arabic, Mandarin Chinese, German, English, Spanish, Hausa, Portuguese, Romanian, Russian, and Ukrainian—and covers a multilingual range of emotion detection. For the full description of the task, dataset, and evaluation setup, refer to the SemEval 2025 Task 11 Track B overview paper (Muhammad et al., 2025b).

Our approach employs a transformer-based model, which depends on multilingual pre-trained language models (PLMs) such as XLM-RoBERTa to acquire semantic and contextualized representations across languages. Because the task is ordinal, we experiment with both regression-based and ordinal classification approaches, complementing data augmentation and fine-tuning methods for improving generalization. We also explore language-specific and multilingual training settings, balancing the trade-offs of cross-lingual knowledge transfer and fine-tuning particular to languages. To further enhance our predictions, we integrate ensemble learning techniques and utilize different model outputs in combination to reduce variance and increase robustness. Through this assignment, we gained valuable lessons in multilingual emotion intensity prediction tasks. Our system performed well with high-resource languages like English and Spanish, performing within the top 60% of submissions. It declined for low-resource languages such as Hausa and Amharic, revealing the limitations of PLMs to handle underrepresented languages. Additionally, our model struggled with subtle distinctions between moderate and high emotion intensities, suggesting directions for future optimization in label calibration. Contrastive learning and emotion-aware embeddings are directions of future work that can enhance the degree of granularity in emotion intensity predictions.

Our code has been publicly released and can be

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1866–1869

accessed at: https://colab.research.google.com/drive/1yDBxSn65gDDzGwZ9trFiHLM6_N7QDXeQ?usp=sharing

2 Background

For Task 11 Track B, the main aim is to create a model that could predict the perceived intensity of emotions within a sentence, across different languages. More specifically, each language had 5-6 major emotions that could be detected in a sentence– joy, sadness, fear, anger, surprise, and disgust. The predictions are a perceived intensity on a scale of 0-3 of each emotion within a sentence, with 0 being no emotion at all and 3 being strong emotion. The datasets provided for this algorithm included the languages Amharic, Algerian Arabic, Mandarin Chinese, German, English, Spanish, Hausa, Portuguese, Romanian, Russian, and Ukrainian. There were separate datasets for dev, test, and train, with varying amounts of data between each language (usually a couple thousand sentences for each dataset).

3 System Overview

Our detection system is built using PyTorch and the HuggingFace transformers library. First for preprocessing, the dataset will be tokenized using the AutoTokenizer from HuggingFace transformers and padded to the default max_length for the model. The labels(emotions) are also converted to tensors for training.

We used the Bidirectional Encoder Representations from Transformers, or BERT, base model (uncased) from Hugging Face as a pre-trained model that we then fine-tuned for a regression task on the provided emotion intensity dataset. We set the tokenizer to the BERT tokenizer, and ran each of the datasets (train, val, test) through our preprocessing pipeline.

For the model itself, we used the BERT model for sequence classification with 5 labels for the emotions (6 for languages with 6 emotions), and set the problem type to regression for this task. We used an AdamW optimizer with a learning rate of 5e-5. In the training loop, we iterated over each batch of data and evaluated the output of the model given tokenized input ids and attention masks that allows the model to differentiate between actual tokens and padding. The loss was then calculated in each iteration using mean squared error, which is the default for a regression task. At the end of the loop, gradients are calculated with loss.backward() and model weights are updated.

4 4 Experimental Setup and Methods

4.1 Data Splits

We use the given dataset for SemEval 2025 Task 11 Track B, dividing it into training, validation, and test sets:

- Training Set: It is used to train the model.
- Validation Set: It is utilized for hyperparameter tuning and early stopping.
- Test Set: It is utilized for final performance evaluation.

The data set consists of text snippets with annotated perceived emotion intensities (anger, disgust, fear, joy, sadness, surprise) on an ordinal scale from 0 to 3. The data is read from CSV files:

- track_b_data/train/[language].csv
- track_b_data/dev/[language].csv
- track_b_data/test/[language].csv

4.2 Preprocessing

Tokenization: We tokenize text data with the Hugging Face AutoTokenizer and the bert-base-uncased model. Sequences are padded or truncated to a maximum of 128 tokens. Dataset Formatting: We define a custom PyTorch Dataset class (EmotionDataset) to handle tokenized input and corresponding labels. Batching: The data is batched with DataLoader for training and testing with batch size 16.

4.3 Model and Training Configuration

Model: BERT-base-uncased is fine-tuned for regression using AutoModelForSequenceClassification with num_labels=1 to return emotion intensity scores. Loss Function: Mean Squared Error (MSE) loss is used to train the regression model. Optimizer: AdamW optimizer with 1e-5 learning rate and weight decay of 0.001. Scheduler: Learning rate is adjusted dynamically with an exponential scheduler (gamma=0.99).

4.4 Training Strategy

Epochs: 15 epochs with early stopping when validation loss does not show improvement for 3 consecutive epochs. Gradient Updates: Backpropagation through loss.backward() and optimization steps through optimizer.step(). Validation Checkpoints: The model is validated at the end of every epoch on the validation set and the best performing model (according to validation loss) is stored.

4.5 Evaluation Metrics

Loss: At validation time, Mean Squared Error (MSE) loss is tracked. Accuracy Proxy: A prediction is considered to be correct if its absolute deviation from ground truth is less than 0.5. Final Output Processing: Prediction is rounded and clamped in the interval [0,3] prior to submission.

4.6 Tools and Libraries

PyTorch (torch==2.x) - Model training and testing. Transformers (transformers==4.x) - Tokenization and model loading. Pandas (pandas==1.x) - Data management and CSV operations.

4.7 Model Saving and Inference

The best model according to validation loss is saved as model_best_weights.pt. Predictions on the test set are saved in [language]_predictions_rounded.csv. This setup ensures reproducibility and effective model training for emotion intensity prediction in the SemEval 2025 Track B task.

5 Results

The proposed model achieved a validation accuracy of 68.97% and a validation loss of 0.373, indicating robust performance in emotion intensity prediction. These results affirm that regression-based fine-tuning of BERT effectively captures subtle variations in emotional expression. Our system ranked within the top 60% overall, demonstrating competitive performance across multiple languages. The model performed well in high-resource languages such as English and Spanish but exhibited limitations in low-resource languages like Hausa and Amharic, suggesting potential constraints in cross-lingual transfer learning.

6 Analysis

6.1 Quantitative Analysis and Ablation Studies

The choice of model architecture played a crucial role in performance. The use of bert-base-uncased provided a strong baseline, but alternative models such as roberta-base and deberta-v3-base could offer enhanced contextual representations. Training on a merged multilingual dataset proved beneficial for high-resource languages, yet it provided limited advantages for low-resource languages, suggesting that improved cross-lingual learning strategies are necessary. The implementation of AdamW with an ExponentialLR scheduler contributed to training stability. Experiments with various batch sizes and learning rates confirmed that our chosen hyperparameters struck an optimal balance between convergence speed and generalization.

6.2 Error Analysis and Limitations

The model encountered challenges in distinguishing moderate from strong emotion intensities, particularly in emotions such as sadness and fear, where contextual subtleties are crucial. False positives and negatives were more frequent in ambiguous cases where multiple emotions co-occurred, indicating the need for more sophisticated emotion-aware embeddings. Overfitting risks were observed, with superior performance on high-resource languages compared to low-resource ones, likely due to dataset imbalances and domain mismatches. While formal human evaluation was not conducted, manual inspection suggested that the model occasionally exhibited bias toward the dominant emotion present in the training data.

7 Conclusion

This study demonstrates that transformer-based models, particularly BERT, are effective for emotion intensity prediction. However, challenges remain in handling low-resource languages and refining distinctions between emotion intensities. Future work will explore alternative architectures such as roberta-base and deberta-v3-base, as well as incorporating contrastive learning techniques to improve representation learning. Expanding dataset diversity is essential to enhance generalization across languages. Furthermore, conducting human evaluations will provide deeper insights into model predictions and refine calibration strategies. Despite these limitations, our approach con-

tributes to the growing field of computational emotion analysis, underscoring the value of pretrained transformers in emotion intensity estimation. The findings provide a strong foundation for future advancements in emotion-aware natural language processing applications.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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, and 29 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv 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, and 2 others. 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.