CharsiuRice at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Hai-Yin Yang **Hing Man Chiu** **Hiu Yan Yip**
Eberhard Karls University of Tübingen
{hai-yin.yang, hing-man.chiu, hiu-yan.yip}@student.uni-tuebingen.de

Abstract

This paper presents our participation in SemEval-2025 Task 11, which focuses on bridging the gap in text-based emotion detection. Our team took part in both Tracks A and B, addressing different aspects of emotion classification. We fine-tuned a RoBERTa base model on the provided dataset in Track A, achieving a Macro-F1 score of 0.7264. For Track B, we built on top of the Track A model by incorporating an additional non-linear layer, in the hope of enhancing Track A model's understanding of emotion detection. Track B model resulted with an average Pearson's R of 0.5658. The results demonstrate the effectiveness of fine-tuning in Track A and the potential improvements from architectural modifications in Track B for emotion intensity detection tasks.

1 Introduction

Emotion is defined as "a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event" (American Psychological Association, 2018). Humans express emotions through speech and behavior, but recognizing and empathizing with emotions is not always straightforward, as emotions are abstract and subjective.

In the digital era, emotion detection is increasingly valuable in applications such as chatbots and AI writing assistants. SemEval-2025 Task 11 (Muhammad et al., 2025a) challenges participants to detect emotions and their intensities from the speaker's perspective across multiple languages. We focused on English due to team familiarity and early dataset availability.

The task includes three tracks. Track A targets multi-label classification of perceived emotions; Track B predicts their intensities; and Track C extends emotion detection to an unseen target language using a model trained only on English. Our submission covers Track A and Track B.

2 Background

2.1 Task and Data Description

We participated in SemEval Task 11 (Muhammad et al., 2025b), addressing both Track A and B of the English tasks. These tracks focus on the classification of multi-label emotions and the quantification of their intensity in English utterance, respectively. Our analysis relied exclusively on datasets provided by SemEval (Muhammad et al., 2025a). As outlined in Table 1, Track A dataset for annotate texts with five primary emotions: anger, fear, joy, sadness, and surprise. The training subset consists of 2,768 instances of short texts, while the development subset contains 116 instances.

Track B consists of 2,768 instances as well, but as detailed in Table 2, the emphasis for this subtask is on quantifying the intensity of emotions. This training subset features a distribution of emotional intensities from 0 (absence of the specific emotion), 1 (least intense), 2 (moderately intense) to 3 (most intense) across various emotions including anger, fear, joy, sadness, and surprise.

2.2 Related Works

The task of multi-label emotion classification has seen various approaches, primarily based on advancements in deep learning technologies. Among these, transformer-based models such as BERT, RoBERTa, and DeBERTa have been widely adopted due to their proficiency in understanding complex language nuances (Devlin et al., 2018). Our method extends these innovations by integrating fine-tuned versions of these models to better suit the specific requirements of emotion classification and intensity prediction.

Several studies have informed our approach. RoBERTa, an optimized version of BERT that demonstrates significant advancements in various

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1082–1088
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Dataset Emotion	Train
Anger	333
Fear	1,611
Joy	674
Sadness	878
Surprise	839
Total	2,768

Table 1: English training dataset distribution for Track A with multi-label emotions.

Intensity Emotion	0	1	2	3
Anger	2,435	207	88	38
Fear	1,157	857	546	208
Joy	2,094	449	161	64
Sadness	1,890	505	248	125
Surprise	1,929	588	215	36
Total	9,505	2,606	1,258	471

Table 2: Track B Training Dataset showing emotion intensities from 0 (least) to 3 (most).

natural language processing tasks, including emotion classification. This model showcases its effectiveness over traditional BERT models due to enhancements in the training procedure and model architecture, making it particularly suited for tasks that require nuanced language understanding (Liu et al., 2019).

3 System Overview

As briefly explained previously, we opted for the BERT models. to cater both Track A (multi-label emotion classification) and Track B (multi-emotion intensity level estimation) tasks.

Given the related nature of Tracks A and B, we adopted a unified training approach to take advantage of semantic overlap between the two tasks. In particular, the model was first fine-tuned on Track A data, which focuses on classifying the presence of emotions inside the text. Afterwards the model training is followed by another fine-tuning on Track B data, which involves estimating the intensity of each emotion. This sequential training allowed the model to build a rich understanding of emotion-related features and semantic information from Track A before further refining its ability to handle emotion intensity nuances in Track B.

3.1 Multi-Label Classification

The BERT model was trained as a straightforward standard sequence classifier without additional modeling or algorithm. The raw text data are first tokenised by the pre-trained Transformer tokeniser, then put into the training loop with the help of Huggingface Transformers library. (Wolf et al., 2020b).

Specifically, we intended to fine-tune the pre-trained BERT model in two sequential phases to optimize its performance for the multi-label emotion classification task. This strategy aimed to maximize the model's ability to map textual inputs to corresponding emotional categories effectively.

In the first phase, the BERT model would be fine-tuned using a publicly available single-label emotion dataset hosted on Hugging Face (Saravia et al., 2018). The dataset covers a large part of the target emotions for this task, including anger, fear, joy, and sadness, without surprise. We hoped that this pre-fine-tuning step would allow the model to familiarize itself with the task of emotion detection, focusing on understanding the relationships between textual inputs and specific emotions in a simplified single-label context. The model's classification head was initially configured to "single label classification" to handle the single-label classification. By pre-fine-tuning with this dataset, the model gained a foundational understanding of emotion detection, preparing it to handle more complex multi-label tasks in subsequent training stages.

Following the pre-fine-tuning stage, the model would be further fine-tuned using the actual true training datasets from Track A. This step further expanded the diversity of the fine-tuning data. For Track A, the model's end-task, the classification head, was reconfigured to "multi-label classification", enabling it to assign multiple emotions to a single text entry.

3.2 Emotion Intensity

The model trained with Track A task were brought to undergo further fine-tuning with Track B objective of predicting emotion intensity levels. We added non-linear layers on top of its output, consisting of a fully connected layer with 128 units and ReLU activation, followed by an output layer for intensity prediction, on a scale of 0 to 3. The base RoBERTa model was unfrozen to allow further fine-tuning on the parameters for the intensity prediction task specifically.

4 Experimental Setup

All experiments were conducted on Google Colab using GPU runtime. The software environment included Python 3.8, PyTorch 1.12.0, and the Hugging Face Transformers library (Wolf et al., 2020a). To ensure reproducibility, we fixed the random seed to 42 and documented all preprocessing steps and hyperparameters.

4.1 Design and Procedure of Track A

In our experimental framework, we evaluated three models from Hugging Face: BERT, RoBERTa, and DeBERTa (He et al., 2021). Preliminary results showed that RoBERTa achieved the highest Micro-F1 score of 61.0 without modifications, outperforming BERT (58.0) and DeBERTa (56.0). Consequently, RoBERTa was selected as the base model for further experimentation. We fine-tuned RoBERTa using the Hugging Face Trainer API and employing the BERT tokeniser.

In the subsequent development phase, our methodology was to utilize all of the training dataset from Track A. During this phrase, hyperparameters were initially selected at random, which led to an improved Micro-F1 score of 0.68. In search of further improvement in model performance, we integrated Optuna for systematic hyperparameter optimization, shown in Figure 1 . Ultimately, our final results were quantified with a Micro-F1 score of 0.7263, reflecting a robust improvement through iterative refinements in our model training and parameter tuning processes.

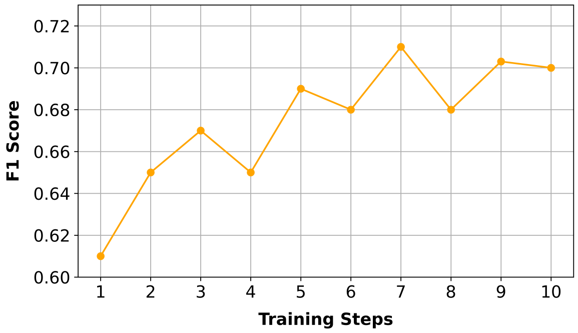

Figure 1: Training Track A's F1 scores over 10 steps, demonstrating the optimization effects of using Optuna. Each point reflects the F1 score adjusted through hyperparameter tuning at each training step, underscoring the effectiveness of the optimization process.

4.2 Design and Procedure of Track B

For training this model, we employed a Mean Squared Error (MSE) loss function, which is particularly suitable for regression tasks. Optimization was carried out using the AdamW optimizer set at a learning rate of $1e-5$. To assess the model's performance, we relied on two primary metrics: the MSE and the Pearson correlation coefficient. Through our training and optimization process, we finally reached a Pearson score of 0.57.

The training regimen involved three full epochs, with the model processing the entire dataset in each epoch. The process entailed executing a forward pass to generate predictions from the input data, followed by the calculation of MSE loss. The model's parameters were then updated via backpropagation to minimize the loss, thereby refining the model's ability to accurately predict emotional intensity. The settings for the training included a batch size of 16.

5 Results and Analysis

This section reports on our models' performances on the test sets of Track A and B, plus analysis of results and system errors. About the official evaluation metrics, Track A uses Marco-F1 score between model prediction and gold labels, while the average Pearson's R over language-specific emotions is the metric of Track B's performance. Jaccard index for Track A and Mean Absolute Error (MAE) for Track B are added for more in-depth understanding of models' performance.

5.1 Track A: Multi-label Emotion Detection

Emotion	**F1 Score**
Anger	0.6132
Fear	0.8286
Joy	0.7538
Sadness	0.7353
Surprise	0.7010
Micro-F1	0.7588
Macro-F1	**0.7264**

Table 3: Individual and aggregated (macro and micro) F1 scores for all 5 English emotions (anger, fear, joy, sadness, surprise).

According to Table 3, our team achieved the Macro-F1 score of 0.7264. It indicates that, on average, our Track A model accurately predicted the presence and absence of the five target emotions

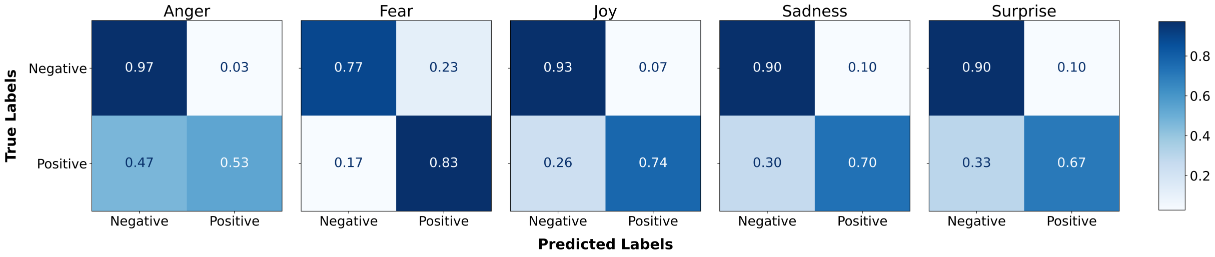

Figure 2: Track A model performance in multi-label emotion detection: comparison of true (y-axis) against predicted labels (x-axis) across 5 emotions using normalized confusion matrices by proportion.

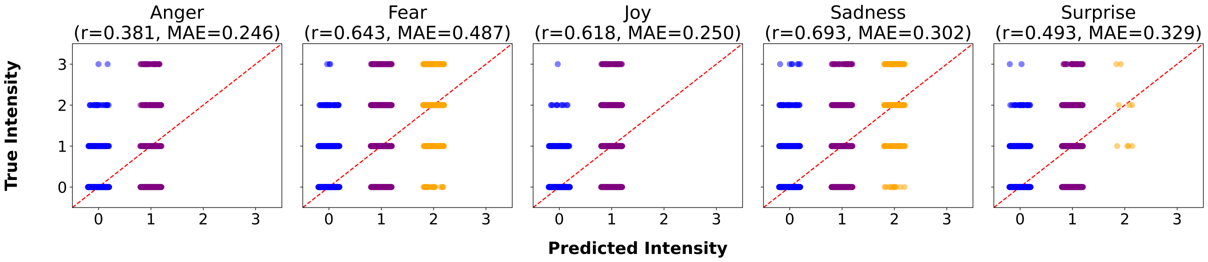

Figure 3: Track B model performance in predicting emotion intensity: comparison of true (y-axis) against predicted intensity (x-axis) across 5 emotions with Pearson's R (r) and Mean Absolute Error (MAE).

72.64% of the time. Among these emotions, the model demonstrated the highest performance in detecting fear, attaining an individual F1 score of 0.8286. Similarly, the detection of joy, sadness, and surprise yielded robust F1 scores of 0.7538, 0.7353, and 0.7010, respectively, all surpassing the 0.7 threshold.

However, the model exhibited noticeable inadequacy in detecting anger, with an F1 score of 0.6132, markedly lower than that of the other four emotions. This score represents the most substantial performance gap among all emotions, with nearly a 0.09-point difference compared to surprise, the emotion with the second-lowest F1 score. The discrepancy between the Macro-F1 and Micro-F1 scores can be attributed to the Macro-F1's sensitivity to rare emotions, such as anger, which significantly impacts the overall average when detection performance is inconsistent.

5.1.1 Individual Emotion Detection Performance

The normalized confusion matrices across the five emotions, presented in Figure 2, provide a more detailed breakdown of the model's performance. Beginning with fear, the model accurately identified 83% of true positive instances while correctly classifying 77% of true negative cases.

In contrast, Figure 2 reveals an opposite trend for the detection of joy, sadness, and surprise. The model demonstrated strong capability in identifying the absence of these three emotions, with over 90% of true negatives correctly classified. Regarding emotion presence, the model successfully detected 74% of joyful instances, 70% of sad instances, and 67% of surprising instances.

Anger stands out as the emotion with the highest true negative detection rate, reaching 97%. However, the model struggled with identifying its presence, falsely classifying 47% of true anger instances as absent. Consequently, only 53% of genuinely angry instances were correctly detected.

5.1.2 Multi-Label Emotion Prediction Accuracy

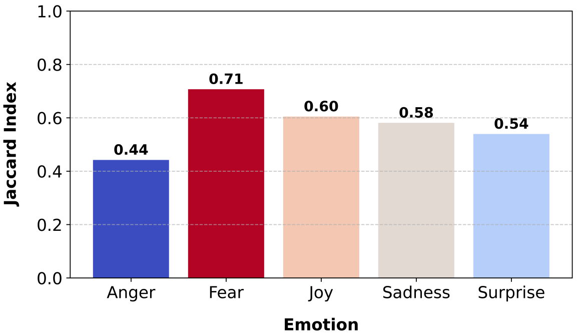

Figure 4: Jaccard index of Track A multi-label emotion detection across 5 emotions.

Figure 4 presents the Jaccard Index scores for each English emotion in Track A, evaluating the

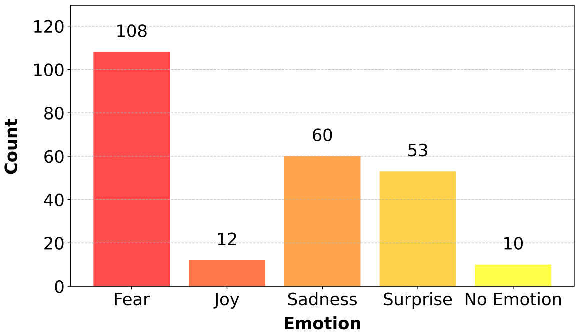

Figure 5: Result distribution of anger's false negative.

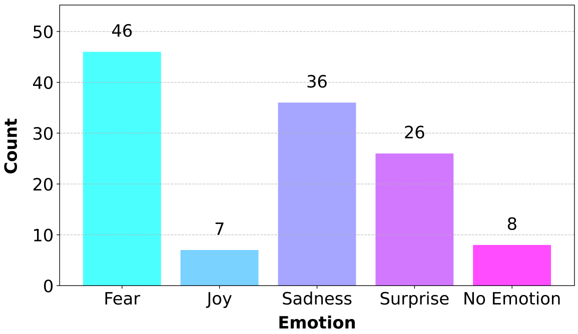

Figure 6: Result distribution of anger's false positive.

intersection-over-union of predicted and actual labels, which gives insights about how well the model predicts the full sets of emotions per instance.

Among the five emotions, the model achieved the highest Jaccard score for fear (0.71), suggesting that the Track B model not only detects fear accurately but also captured other emotions that typically occur with it. Similarly, joy (0.60) and sadness (0.58) yielded moderate Jaccard scores, indicating that while the model correctly predicts these emotions in multi-label contexts, some mismatches still occur.

However, the model struggled the most with anger (0.44) and surprise (0.54). Barely half of angry instances were detected correctly. We will dive into anger as the most struggling emotion to be detected below.

Figures 5 and 6 illustrate the misclassification patterns of anger predictions by the model, comparing false negatives and false positives. The left graph shows cases where the model missed detecting anger and instead predicted other emotions. Notably, fear (108 cases) was the most frequent misclassification, followed by sadness (60 cases) and surprise (53 cases), indicating that anger is often confused with emotionally intense states. In contrast, joy (12 cases) and no emotion (10 cases) were rarely chosen, suggesting that the model differentiates them well from anger. The right graph presents false positives, where the model incorrectly predicted anger instead of the actual emotion. Fear (46 cases) and sadness (36 cases) were the most frequent true emotions mislabeled as anger, with surprise (26 cases) also contributing significantly. This pattern suggests that the model overpredicts anger in contexts where strong emotional expressions are present, particularly in fearful or sad statements.

5.1.3 Team Ranking of English Track A

Our team ranked as the 44th among 95 teams with the Macro-F1 score of 0.7264. It is 0.018 higher than the baseline model, while the best performed model has achieved the Macro-F1 score of 0.8230.

5.2 Track B: Emotion Intensity

Emotion	Pearson's R
Anger	0.3813
Fear	0.6434
Joy	0.6178
Sadness	0.6932
Surprise	0.4933
Average Pearson's R	**0.5658**

Table 4: Individual and average Pearson's R scores for all five English emotions.

According to Table 4, our team achieved an average Pearson's R of 0.5658 in identifying emotion intensity, indicating a moderate correlation between the gold-standard and predicted intensity labels.

Among the five emotions, the model performed best in predicting the intensity of sadness, achieving a Pearson's R of 0.6932. Comparable performance was observed for fear and joy, with correlation coefficients of 0.6434 and 0.6178, respectively. The model exhibited moderate performance in predicting the intensity of surprise, with a Pearson's R of 0.4933.

However, the model faced again notable challenges in predicting the intensity of anger, attaining a Pearson's R of only 0.3813. This highlights a significant performance gap compared to other emotions, emphasizing the need for further refinements to better recognise and predict nuanced variations in anger intensity.

5.2.1 Prediction Correlation Trend

The jittered strip plots in Figure 3 provide further insight into our Track B model's performance. The red dashed diagonal lines represent perfect predictions ($Y_{true} = Y_{pred}$), where predicted values align exactly with the gold-standard intensities. Dots above the dashed line indicate underestimation, while dots below the dashed line indicate overestimation.

Among all emotions, fear and sadness were the most accurately predicted, with most intensity values ranging from 0 to 2 and a considerable number of dots lying on the dashed line. Similarly, joy was well predicted, but the model's predictions were mostly confined to the range of 0 to 1. For surprise, predictions were concentrated between 0 and 1, with the model rarely predicting an intensity of 2. The weakest alignment was observed in anger, where the model's predictions were mostly limited to 0 and 1, failing to capture higher intensity levels.

5.2.2 Prediction Accuracy: Mean Absolute Error (MAE)

Beyond correlation, we further assess Track B model's predictive accuracy using Mean Absolute Error (MAE), which measures the average absolute difference between predicted and true intensity values. While Pearson's R evaluates trend-following ability. MAE provides a more direct measure of prediction accuracy.

As noted in Figure 3, across all five emotions, the lowest MAE was observed for anger (0.246) and joy (0.250), indicating that the model's predictions for these emotions were generally close to the true values. However, as reflected by the low Pearson's R for anger (0.3813), this low MAE primarily results from the model consistently predicting within a limited range (0-1), failing to capture higher intensity variations.

In contrast, the highest MAE was observed for fear (0.487). This aligns with the trend in Figure 3, where several extreme errors, such as cases of predicting 0 when the true intensity was 2 or 3, were observed. The model also exhibited a relatively high MAE for surprise (0.329), largely due to its tendency to underestimate high-intensity instances, as seen in the absence of predictions at intensity levels 2 and 3.

The model's performance on sadness (0.302) represents a balance between strong correlation ($r = 0.6932$) and moderate MAE, indicating that while the model successfully captures the general trend of sadness intensity, it still exhibits noticeable absolute errors in individual predictions. This reflects an important insight: high Pearson's R does not necessarily equate to low MAE, as the model may effectively follow intensity ranking trends while making significant absolute magnitude errors.

5.2.3 Team Ranking of English Track B

Our model did not perform better than the baseline model, which has an average Person's R gap of 0.08. We rank at the place of 37 out of 43 teams.

5.3 Future Enhancement

In both Track A and B, our models struggle to detect anger and high intensity of emotion. It is mainly due to the reason of imbalanced training data. As depicted in Table 1 and 2, comparing to 1,611 fearful instances in Track A provided training data, fear has only 333 instances. For Track B the highest intensity (3) has only 471 examples, while level 0 has 9,505 instances for the model to learn.

To address this issue, we could try boosting the training data, such as data augmentation to artificially generate data that training set does not cover much. Data balancing measures are also crucial to reduce model's bias towards specific classes.

Going further, other training techniques such as transfer learning or ensemble learning, or taking multiple machine learning algorithms on these classification and regression tasks can also be considered for potential experiments.

6 Conclusion

We participated in Tracks A and B of the shared task by fine-tuning BERT-based models for multi-label emotion detection and emotion intensity prediction on English texts. Our model ranked 44th in Track A—around the median—but underperformed compared to the baseline in Track B.

The results highlight the challenges of emotion detection, even in a high-resource language, as emotions are often implicit and not directly conveyed through surface-level text. They also suggest that full fine-tuning may not be optimal given the dataset size and distribution.

We hope our work offers insights for future research and the development of emotion-aware applications using pre-trained language models.

References

American Psychological Association. 2018. Emotion - APA dictionary of psychology.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

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.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. 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. 2020a. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020b. Huggingface's transformers: State-of-the-art natural language processing.