NLP-DU at SemEval-2025 Task 11: Analyzing Multi-label Emotion Detection

Md. Sadman Sakib[1], **Ahaj Mahhin Faiak**[1]

Abdullah Ibne Hanif Arean[1], **Fariha Anjum Shifa**[1]

[1]Department of Computer Science and Engineering, University of Dhaka

{mdsadman-2020015659@cs.du.ac.bd, farihaanjum.24csedu.022@gmail.com}

{abdullaharean2613@gmail.com}

Abstract

This paper describes NLP-DU's entry to SemEval-2025 Task 11 on multi-label emotion detection. We investigated the efficacy of transformer-based models and propose an ensemble approach that combines multiple models. Our experiments demonstrate that the ensemble outperforms individual models under the dataset constraints, yielding superior performance on key evaluation metrics. These findings underscore the potential of ensemble techniques in enhancing multi-label emotion detection and contribute to the broader understanding of emotion analysis in natural language processing.

1 Introduction

Emotion detection seeks to identify and categorize emotions conveyed in textual data. This task presents significant challenges due to the inherently complex and overlapping nature of human emotions, as well as the difficulty associated with acquiring high-quality labeled datasets. Muhammad et al. (2025a) introduce the BRIGHTER dataset, which aims to bridge the gaps in textual emotion recognition across 28 languages and further enriched their semantic evaluation work in Muhammad et al. (2025b). Furthermore Belay et al. (2025) introduce Evaluating the Capabilities of Large Language Models for Multi-label Emotion Understanding. Recent research by Zhang et al. (2020) underscores the necessity of modeling both label dependence and modality dependence in multi-modal, multi-label emotion detection, further highlighting the intricacies involved in this domain.

In our approach, we first explored CNN and LSTM-based solutions and checked for baseline performance after training. Then we chose to explore transformer-based models for multi-label emotion detection. We experimented with ModernBERT Warner et al. (2024), DeBERTa He et al. (2021), ALBERT Lan et al. (2020), XLM-RoBERTa Conneau et al. (2020), DistilBERT Sanh et al. (2020), and XLNet Yang et al. (2020), some state-of-the-art transformer models known for their strong contextual understanding and generalization capabilities. To address the challenges posed by a compact dataset Muhammad et al. (2025a) with sensitive labeling and emotion context added to each sentence, we employed data augmentation techniques and leveraged multiple dataset splitting techniques such as balanced stratified K-fold splitting, tree-based splitting, k-means splitting, and balanced stratified splitting for robust evaluation. Additionally, we propose an ensemble approach that combines multiple transformer-based models, for example, ModernBERT and DeBERTa, via weighted averaging, improving overall performance. To encourage reproducibility, we have released our code and models, which can be accessed at: https://github.com/ssadman887/SEMEVAL-TASK-2025.

2 System Overview

2.1 Data Description

The dataset consists of text samples labeled with multiple emotional categories: Anger, Fear, Joy, Sadness, and Surprise. The analysis of text length distribution reveals that most samples contain between 5 to 25 words, with a right-skewed distribution indicating that shorter texts are more common. In terms of class distribution, the dataset is imbalanced, with Fear and Anger appearing more frequently than Surprise and Joy.

The correlation analysis between emotions shows notable relationships, such as a strong negative correlation between Joy and Fear (-0.49), while Fear and Sadness exhibit a moderate positive correlation (0.27). These findings provide insights into the dataset's structure, which is essential for guiding preprocessing steps and model training.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1269–1275

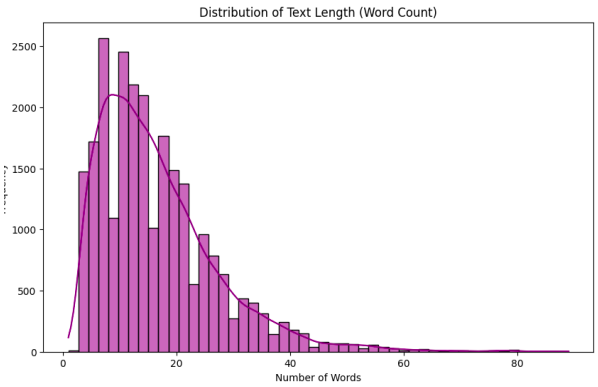

Figure 1: Distribution of Text Length

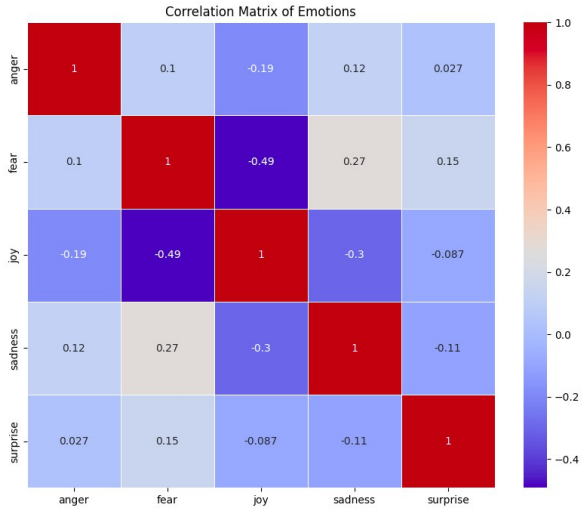

Figure 2: Correlation Matrix of Emotions

2.2 Key Algorithms and Modeling Decisions

Our system for multi-label emotion detection is based on a transformer-based architecture. The system follows a sequence classification approach where each input text is encoded into contextual embeddings and passed through a multi-label classification head.

2.2.1 Data Augmentation

To improve dataset diversity and robustness, a sentence rewriting data augmentation strategy was utilized. This method rephrased sentences while preserving original emotion labels. By creating various versions of the same text, the dataset expanded, enhancing the model's capacity to generalize across diverse linguistic emotion expressions. This technique increased training data volume and introduced more syntactic and lexical variety, thereby boosting the model's ability to recognize nuanced emotional expressions. We used Meta LLaMA 3.1

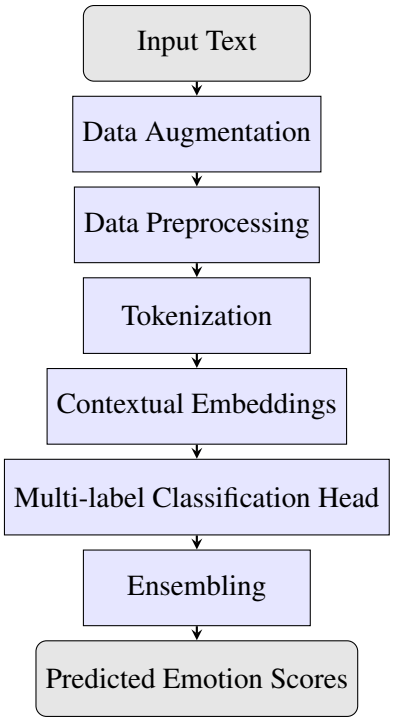

Figure 3: Pipeline for multi-label emotion classification

AI (2025) from the Ollama platform on a local device using an RTX 4050 GPU. We employed multiple augmentation techniques, discussed in Table 1. We checked each data row for discrepancies and modified accordingly.

2.2.2 Preprocessing Steps

The preprocessing phase aimed to preserve the data's textual integrity while aligning it with model architectures. Minimal text cleaning was conducted to maintain the original semantic meaning. This method retained critical linguistic structures, avoiding the unintended loss of emotional nuances. Next, tokenization was performed using the tokenizers specific to ModernBERT, DistilBERT, and DeBERTa, capping sequence length at 512 tokens. This converted raw text into structured tokens for transformer-based model processing. Lastly, label representation used binary encoding for each emotion category, enabling the model to predict multiple emotions per input, thus reflecting the complex interdependencies of human emotional expressions.

2.3 Model Architecture

The model architecture consists of ModernBERT, DistilBERT, and DeBERTa as the base models. A classification head is applied on top, which includes a fully connected layer with a 0.01 dropout rate to

Augmentation Technique	Original Sentence	Augmented Sentence
Synonym Replacement	They were dancing to Bolero.	They were performing to Bolero.
Perspective Transformation	I moved my arms, stretching the muscles.	He moved his arms, stretching the muscles.
Voice Transformation	The cop tells him to have a nice day.	He was told to have a nice day by the cop.
Tone Adjustment	We ordered some food at McDonald's instead of buying food at the theatre because of the ridiculous prices the theatre has.	We opted for McDonald's rather than purchasing food at the theatre due to its exorbitant prices.
Tense Consistency	About 2 weeks ago I thought I pulled a muscle in my calf.	About 2 weeks ago I had thought I had pulled a muscle in my calf.
Tag Question Addition	The room was small but brightly lit.	The room was small but brightly lit, wasn't it?
Neutral Modifier Insertion	I still cannot explain this.	I still cannot quite explain this.

Table 1: Examples of Data Augmentation Techniques Applied Using Meta LLaMA

prevent overfitting. The output layer uses a sigmoid activation function to predict multi-label probabilities. The model is trained using the Binary Cross Entropy with Logits loss function, optimized with AdamW at a learning rate of 2e-5.

2.4 Training Strategy

The training strategy was designed to optimize model performance for multi-label emotion classification while ensuring robustness and generalization across diverse data subsets. We employed a 5-fold Multilabel Stratified Cross-Validation (MSCV) approach, which preserves the label distribution across folds in a multi-label setting. For a dataset D with N samples and $K = 5$ labels (Anger, Fear, Joy, Sadness, Surprise), MSCV partitions D into 5 folds $\{D_1, D_2, \ldots, D_5\}$, where each fold D_i maintains the proportion of positive instances for each label k:

$$\text{prop}_{k,i} \approx \frac{1}{N} \sum_{n=1}^{N} y_{n,k}, \quad \forall i \in \{1, 2, \ldots, 5\}, \quad (1)$$

where $y_{n,k} \in \{0, 1\}$ is the k-th label for the n-th sample, and $\text{prop}_{k,i}$ is the proportion of positive instances for label k in fold D_i. This stratification ensures that the model is trained and evaluated on representative subsets, mitigating bias due to label imbalance.

To prevent overfitting and optimize training efficiency, early stopping was applied based on the change in macro F1 score between epochs. Let $\text{F1}^{(e)}_{\text{macro}}$ denote the macro F1 score on the validation set at epoch e. Early stopping was triggered if the improvement in F1 score was below a threshold:

$$\Delta\text{F1} = \text{F1}^{(e)}_{\text{macro}} - \text{F1}^{(e-1)}_{\text{macro}} < 0.01, \quad (2)$$

halting training to retain the model weights from the epoch with the highest validation performance. Table 2 summarizes the training hyperparameters.

Training Parameter	Value
Batch Size	16
Epochs	5
Early Stopping Threshold	$\Delta\text{F1} < 0.01$
Optimizer	AdamW
Learning Rate	2×10^{-5}

Table 2: Training hyperparameters for model optimization.

2.5 Ensembling Strategy

To enhance prediction robustness, we employed Weighted Ensembling, where predictions were combined based on model confidence:

$$\hat{y}_{\text{ensemble}} = w_A \cdot \hat{y}_A + w_B \cdot \hat{y}_B \quad (3)$$

where $w_A, w_B \in [0, 1]$ and typically $w_A + w_B = 1$ to ensure normalized weighting. Furthermore, we tested both the best model among the folds and, in other cases, used all folds to predict results. In a multi-label classification setting with M models and K labels, majority voting aggregates binary predictions to produce a consensus prediction. For the n-th sample and k-th label, let $\hat{y}_{m,n,k} \in \{0, 1\}$ represent the binary prediction

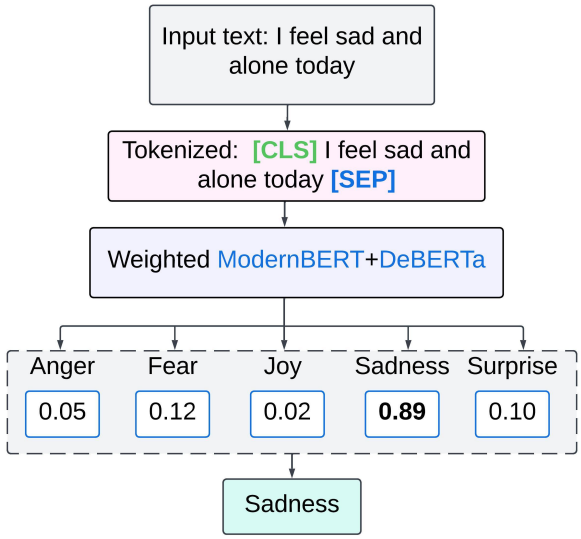

Figure 4: An example of emotion classification of an input text

from the m-th model. The ensemble prediction $\hat{y}_{n,k}$ is determined by majority voting:

$$\hat{y}_{n,k} = \mathbb{1}\left(\sum_{m=1}^{M} \hat{y}_{m,n,k} \geq \left\lceil \frac{M}{2} \right\rceil\right), \quad (4)$$

where $\mathbb{1}(\cdot)$ is the indicator function, returning 1 if the condition is true and 0 otherwise, and $\lceil x \rceil$ denotes the ceiling function.

2.6 Example Walkthrough

Figure 4 shows an example of emotion classification of an input text. We have taken a sample data from our dataset to show the workflow.

2.7 Addressing Key Challenges

To mitigate the effects of class imbalance, particularly for underrepresented emotions such as Surprise, we implemented a weighted loss function. In a multi-label classification setting, the weighted binary cross-entropy loss adjusts for class imbalance across labels. For the n-th sample with K labels, the loss $\mathcal{L}_n$ is defined as:

$$\mathcal{L}_n = -\frac{1}{K}\sum_{k=1}^{K} w_k \cdot \big[y_{n,k} \cdot \log(\sigma(\hat{z}_{n,k})) + (1 - y_{n,k}) \cdot \log(1 - \sigma(\hat{z}_{n,k}))\big], \quad (5)$$

where $w_k \geq 0$ is the weight for the k-th label, $y_{n,k} \in \{0, 1\}$ is the ground truth, and $\sigma(\hat{z}_{n,k})$ is the sigmoid activation of the logit $\hat{z}_{n,k}$. The total loss over a batch of N samples is:

$$\mathcal{L}_{\text{batch}} = -\frac{1}{N \cdot K}\sum_{n=1}^{N}\sum_{k=1}^{K} w_k \cdot \big[y_{n,k} \cdot \log(\sigma(\hat{z}_{n,k})) + (1 - y_{n,k}) \cdot \log(1 - \sigma(\hat{z}_{n,k}))\big]. \quad (6)$$

To address class imbalance, oversampling balances the dataset by increasing the representation of underrepresented labels. For a dataset D with N samples and K labels, let $y_{n,k} \in \{0, 1\}$ denote the k-th label for the n-th sample. The frequency of positive instances for the k-th label is:

$$f_k = \sum_{n=1}^{N} y_{n,k}. \quad (7)$$

The maximum frequency across all labels is $f_{\text{max}} = \max_k\{f_k\}$. The oversampling ratio for the k-th label is:

$$r_k = \frac{f_{\text{max}}}{f_k}. \quad (8)$$

For each sample (x_n, y_n) where $y_{n,k} = 1$, approximately $\lfloor r_k \rfloor$ duplicates are created to balance the dataset.

3 Experimental Setup

3.1 Data Preparation

We applied data augmentation by altering sentence structures while preserving the original emotions, increasing dataset diversity. Standard data cleaning techniques, including special character removal, and tokenization, were used for preprocessing.

3.2 Data Splitting Strategy

To identify the optimal data partitioning method, we evaluated four strategies. Tree-based splitting utilized hierarchical clustering to group similar data points prior to division. K-Means clustering generated diverse data clusters for balanced splits. Balanced splitting preserved label distribution across partitions. Finally, Multilabel Stratified K-Fold Cross-Validation with five folds maintained label consistency in each fold. These methods were assessed for both training and prediction to determine the most effective solution.

3.3 Training Procedure

Models were trained using a batch size of 16 for up to 5 epochs, with early stopping applied if the improvement in F1 score was below 0.01. The optimizer used was AdamW with a learning rate of 2×10^{-5}.

3.4 Hardware and Software

All experiments were conducted on the Kaggle platform using a GPU P100 and a local Ollama platform using an RTX 4050. The implementation was done using PyTorch, Transformers (Hugging Face), and Scikit-learn.

4 Experimental Results and Analysis

We evaluated both individual transformer-based models and ensemble strategies, using the development (Dev) and test datasets. Performance is assessed through F1 scores (macro and micro) and accuracy, providing a comprehensive view of model effectiveness in this multi-label setting.

4.1 Individual Model Performance

Table 3 summarizes individual model performance. ModernBERT leads on the Dev

dataset with an F1 score of 0.7842 and accuracy of 85.65%, outperforming DistilBERT (F1 = 0.6970, accuracy = 80.17%), XLM-R_{Base} (F1 = 0.6470, accuracy = 77.70%), and XLNet (F1 = 0.6140, accuracy = 71.9%). DeBERTa follows with an F1 score of 0.7324 and accuracy of 84.48%. On the Test dataset, ModernBERT and DeBERTa achieve macro F1 scores of 0.6805 and 0.6930, and micro F1 scores of 0.7212 and 0.7232, respectively. Conversely, XLNet, XLM-R_{Large} (Dev F1 = 0.6342, Test macro F1 = 0.5762), and ALBERT underperform, likely due to pretraining misalignment with emotional text. The LSTM baseline (Dev F1 = 0.4278, Test macro F1 = 0.3805) highlights transformers' superiority.

4.2 Ensemble Model Performance

Ensemble methods enhance performance by combining model predictions. The weighted averaging strategy combines logits as:

$$\hat{z}_{n,k}^{\text{ensemble}} = w_A \cdot \hat{z}_{n,k}^{A} + w_B \cdot \hat{z}_{n,k}^{B}, \qquad (9)$$

with $w_A + w_B = 1$, yielding a prediction $\hat{y}_{n,k} = \sigma(\hat{z}_{n,k}^{\text{ensemble}})$. For ModernBERT+DeBERTa ($w = 0.5, 0.5$), this achieves the highest Dev F1 score (0.7886) and Test macro F1 score (0.7086). Majority voting, defined as:

$$\hat{y}_{n,k}^{\text{majority}} = \mathbb{1}\left(\sum_{m=1}^{M} \hat{y}_{m,n,k} \geq \left\lceil \frac{M}{2} \right\rceil\right), \qquad (10)$$

reaches a Test micro F1 of 0.7457. The best-fold weighted ensemble, with weights:

$$w_m = \frac{\text{F1}_m^{\text{best-fold}}}{\sum_{m'=1}^{M} \text{F1}_{m'}^{\text{best-fold}}}, \qquad (11)$$

achieves the highest Test micro F1 (0.7467), while best-fold voting yields 0.7432. DeBERTa+DistilBERT ensembles underperform (e.g., weighted Dev F1 = 0.7261, Test micro F1 = 0.7418 for best-fold weighted) due to DistilBERT's lower capacity.

4.3 Result Table

Table 3 provides a detailed comparison of all models and ensembles, highlighting the superiority of the ModernBERT+DeBERTa combinations across most metrics.

4.4 Discussion

Transformer-based models outperform traditional architectures like LSTM, with ModernBERT (Dev F1 = 0.7842, accuracy = 85.65%) and DeBERTa (Dev F1 = 0.7324, Test macro F1 = 0.6930) leading due to their advanced pretraining and attention mechanisms. Ensemble methods enhance performance, with ModernBERT+DeBERTa weighted averaging (Equation 3) achieving the highest Dev F1 (0.7886) and Test macro F1 (0.7086), while majority voting (Equation 4) excels in Test micro F1 (0.7457). The best-fold weighted ensemble (Equation 5) yields the top Test micro F1 (0.7467). DeBERTa+DistilBERT ensembles underperform due to DistilBERT's lower capacity (Dev F1 = 0.6970). The poor performance of XLNet, XLM-R_{Large}, and ALBERT suggests their pretraining may not suit emotional text. These results highlight the efficacy of ensemble learning for multi-label emotion classification, with future work potentially exploring dynamic weighting or analyzing underperforming models for architectural improvements.

5 Conclusion

We employed various data splitting techniques and augmentation strategies to enhance the robustness of our training process. One of the key challenges

Model Name	Dev Data		Test Data	
	F1 Score	Accuracy	F1 Score (Macro)	F1 Score (Micro)
LSTM	0.4278	64.20%	0.3805	0.4022
DistilBERT	0.6970	80.17%	0.6521	0.7001
ModernBERT	0.7842	85.65%	0.6805	0.7212
DeBERTa	0.7324	84.48%	0.6930	0.7232
XLM-R_{Base}	0.6470	77.70%	0.5804	0.6234
XLM-R_{Large}	0.6342	76.33%	0.5762	0.6300
XLNet	0.6140	71.90%	0.5742	0.6265
ALBERT	0.5872	72.20%	0.5659	0.6282
Ensemble Models				
ModernBERT+DeBERTa$_{0.5+0.5\ weight}$	0.7886	85.30%	0.7086	0.7443
ModernBERT+DeBERTa$_{majority\ voting}$	0.7639	85.42%	0.7034	0.7457
ModernBERT+DeBERTa$_{best\ fold\ weighted}$	0.7399	86.20%	0.7056	0.7467
ModernBERT+DeBERTa$_{best\ fold\ voting}$	0.7528	86.90%	0.7044	0.7432
DeBERTa+DistilBERT$_{0.5+0.5\ weight}$	0.7261	85.17%	0.7004	0.7322
DeBERTa+DistilBERT$_{majority\ voting}$	0.7128	84.31%	0.6947	0.7212
DeBERTa+DistilBERT$_{best\ fold\ weighted}$	0.7318	85.00%	0.6925	0.7418
DeBERTa+DistilBERT$_{best\ fold\ voting}$	0.7324	84.48%	0.6943	0.7422

Table 3: Performance of individual and ensemble models on Dev and Test datasets. The best ensemble performance is underlined.

we encountered was the inherent imbalance in the dataset, with certain emotions being overrepresented. Additionally, we identified instances of mislabeling within the training data, which introduced noise into the learning process. Despite these challenges, our approach enabled us to achieve a competitive F1 score, demonstrating the effectiveness of our data handling strategies and model optimization techniques.

6 Ethical Considerations

Due to the sensitivity of the training data, there is a risk of misclassification in sentence predictions, potentially leading to incorrect label assignments. This is particularly concerning for emotionally charged content, where misinterpretation could have significant implications. To mitigate such risks, we implemented data augmentation and rigorous evaluation strategies to enhance model robustness. Additionally, we emphasize the importance of human oversight in critical applications to ensure ethical and responsible deployment of our system.

References

Meta AI. 2025. Llama 3.1 8b model. https://huggingface.co/meta-llama/Llama-3.1-8B. Accessed: 2025-02-28.

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. *Preprint*, arXiv:1911.02116.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

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. *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 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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3584–3593, Online. Association for Computational Linguistics.