Exploration Lab IITK at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Tafazzul Nadeem* **Riyansha Singh*** **Suyamoon Pathak*** **Ashutosh Modi**
Indian Institute of Technology Kanpur (IIT Kanpur)
Kanpur, India
{tafazzul23, riyansha, suyamoonp24, ashutoshm}@cse.iitk.ac.in

Abstract

This paper presents our approach to SemEval-2025 Task 11 (Track A): Bridging the Gap in Text-Based Emotion Detection, focusing on multilabel emotion classification for the English dataset. Our methodology leverages an ensemble of transformer-based models, incorporating full fine-tuning along with additional classification layers to enhance predictive performance. Through extensive experimentation, we demonstrate that fine-tuning significantly improves emotion classification accuracy compared to baseline models. In addition, we provide an in-depth analysis of the dataset, highlighting key patterns and challenges. The study also evaluates the impact of ensemble modeling on performance, demonstrating its effectiveness in capturing nuanced emotional expressions. Finally, we outline potential directions for further refinement and domain-specific adaptations to enhance model robustness. Our submission was officially ranked 34th in Track A (multilabel emotion detection) leaderboard for English language.

1 Introduction

The increasing availability of electronic documents in the digital era has provided a valuable resource for analyzing human expressions and improving various applications. Understanding this plays a crucial role in multiple domains, including user experience enhancement, social media analysis, mental health monitoring, and market research. Increasing scholarly attention has been directed toward extracting user perspectives on various events by analyzing textual content. The computational identification and classification of opinions within text has been recognized as a fundamental step in data mining. Researchers have traditionally focused on determining whether a given text conveys a positive, negative, or neutral stance toward a specific subject or product (Feng et al., 2021). More recently, studies have expanded to incorporate multidimensional emotional annotations (Hu and Flaxman, 2018; Tasmin, 2018; Acheampong et al., 2020) capturing sentiments such as joy, fear, anger, etc.

Emotion expression in language is inherently nuanced and complex, presenting challenges for emotion recognition. Despite its significance, research in this field has predominantly focused on high-resource languages, leading to significant disparities in dataset availability and model performance for low-resource languages. To address this gap, the organizers of SemEval-2025 Task 11 have curated a specialized dataset, BRIGHTER (Muhammad et al., 2025a) to bridge further the gap in emotion recognition research in underrepresented languages, facilitating more inclusive and effective NLP models. It is a collection of multilabel emotion-annotated datasets spanning 28 languages. This dataset emphasizes low-resource languages from Africa, Asia, Eastern Europe, and Latin America, incorporating diverse textual sources annotated by fluent speakers. The shared task consists of three tracks: multi-label emotion classification (track A), emotion intensity prediction (track B), and text cross-lingual emotion detection (track C).

In this paper, we present our system developed for track A of the task (Muhammad et al., 2025b), multilabel emotion classification. Transformers have been proven to be the most successful in understanding the contextual and semantic information of the text (Acheampong et al., 2021; Gillioz et al., 2020). We selected top-performing models from preliminary experiments, including cardiffnlp/twitter-roberta-large-emotion-latest (Antypas et al., 2023), SamLowe/roberta-base-goemotions (Lowe, 2022) and Emanuel/twitter-emotion-deberta-v3-base (Huber, 2021), to build an ensemble leveraging these models, combining their strengths for improved emotion classification,

* All Authors have equal contribution.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1859–1865

and performed full fine-tuning. Since these models had more emotion categories than our dataset, we incorporated additional classifier layers, adapting to our task.

Our system demonstrated good performance across track A. We achieved competitive results, ranking 42 on the English dataset among the 75 participating teams. Dataset imbalance presents a challenge in these methods. Robust methods using undersampling and oversampling techniques need to be used to overcome this challenge. Further, extensive data analysis is always beneficial. Please find the code here- https://github.com/Tafazzul-Nadeem/TBED-CS779-IITK.

2 Background

The landscape of human emotions has been described through various taxonomies and frameworks. Ekman (Ekman and Friesen, 1971) identified six core emotions expressed through facial expressions that are universally recognized across cultures: joy, sadness, anger, surprise, disgust, and fear. Later, finer-grained taxonomies have been developed, capturing a broader range of up to 600 emotions and employing machine learning to cluster emotion concepts (Cowen and Keltner, 2019). These frameworks highlight the complex, culturally influenced nature of emotions expressed through vocalization (Cowen and Keltner, 2018), music, and facial expressions.

Previous approaches to text-based emotion detection have primarily utilized machine learning (ML) techniques. For instance, Wikarsa et al. and Ameer et al. (Ameer et al., 2021) focused on multi-label emotion classification for code-mixed SMS messages in Roman Urdu and English, employing classical ML methods (SVM, J48, Naive Bayes, etc.) and deep learning models (LSTM, CNN, etc.) on a new dataset. Their results indicated that classical ML methods outperformed both ML and deep learning models. Similarly, Polignano et al. (Polignano et al., 2020) developed a model combining Bi-LSTM, Self-Attention, and CNN for emotion detection, finding that word embeddings significantly improved performance. Their experiments in the ISEAR, SemEval-2018 (Mohammad et al., 2018), and SemEval-2019 datasets demonstrated that the ISEAR dataset produced the best precision and recall.

In recent years, research on text-based emotion detection has increasingly utilized transformer-based pre-trained language models. For instance, Acheampong et al. (Acheampong et al., 2020) conducted comparative analyses of models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019) for emotion recognition using the ISEAR dataset. Asalah et al. (Thiab et al., 2024) proposed an ensemble deep learning approach for emotion detection in textual conversations, CNN-based model and transformer-based models, including BERT, RoBERTa, and XLNet. They utilize hard and weighted majority voting methods to enhance prediction accuracy. Their method demonstrates superior performance, achieving a micro-averaged F1-score of 77.07% on the SemEval-2019 Task 3: EmoContext dataset (Chatterjee et al., 2019), outperforming previous baseline results.

We performed experiments with the track A dataset (Muhammad et al., 2025a), where the training set contains 2768 samples with five binary emotion labels (joy, sadness, fear, anger, and surprise). Dev set contains 116 samples, and 2767 samples are available for inference.

3 Analysis

The co-occurrence matrix of the labels for the English dataset is plotted to gain insights about the correlation between the labels, as shown in Figure 1. We have also visualized the box plot for

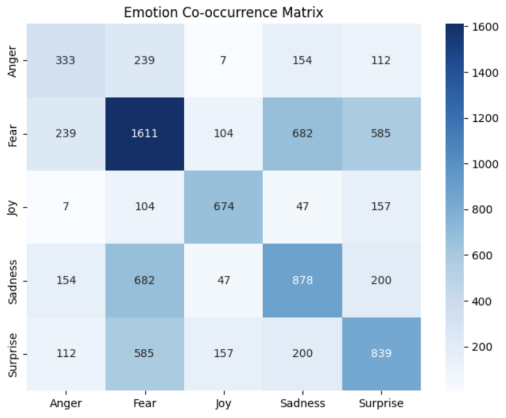

Figure 1: Co-occurrence matrix of emotion labels

length of the text snippets showing the interquartile range (Q1, Q2, etc.) in Figure 2 to gain insights about the context length or prompt length we require while selecting a transformer model. The maximum length is found to be 94 words. Hence, any model with a 512 token size can be used since

the number of tokens = 4 x the number of words (widely used estimate).

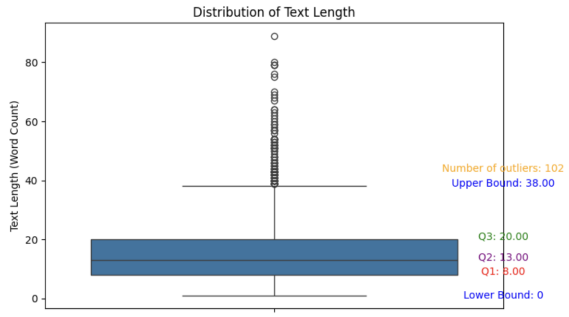

Figure 2: Boxplot of text snippet length

4 Proposed Approach

Since transformer-based models have shown promising results in classification tasks lately, we decided to first find some suitable models for the multilabel classification task. From recent works (Lowe, 2022; Barbieri et al., 2022; Antypas et al., 2023; Huber, 2021; Grattafiori et al., 2024), we shortlisted the following huggingface models, ensembled them, and fully fine-tuned the ensemble: cardiffnlp/twitter-roberta-large-emotion-latest (Antypas et al., 2023), SamLowe/roberta-base-go_emotions (Lowe, 2022) and Emanuel/twitter-emotion-deberta-v3-base (Huber, 2021). The ensemble model, along with the standalone parts, was evaluated on the training data for the English language with full fine-tuning. The results are shown in Table 1.

5 Experiments

The following sections provide a comprehensive overview of our model development process and the final model used for multi-label emotion detection. The first section traces the evolution of our approach, detailing the various models and fine-tuning strategies we experimented with. This includes trials with both smaller and larger models, full fine-tuning, classifier adaptations, and specialized training settings such as entailment-based approaches. The second section focuses on our final approach, which represents the culmination of our iterative experimentation. We describe its architecture, training methodology, and the rationale behind selecting this configuration as our best performing system. The final model integrates insights gained from our earlier experiments, leveraging an ensemble of fine-tuned models to achieve optimal performance.

5.1 Evolution of our approach

We selected some of the best performing models from preliminary experimentation, cardiffnlp/twitter-roberta-large-emotion-latest (Antypas et al., 2023), SamLowe/roberta-base-go_emotions (Lowe, 2022), and Emanuel/twitter-emotion-deberta-v3-base (Huber, 2021), and evaluated them on the validation set to get baseline results. Then, full fine-tuning was performed on these models.

We then tried training adapter models by adding a classifier layer. Since the off-the-shelf models have more emotion categories than our dataset, we added an extra fully connected (FC) layer with five neurons (equal to emotion labels in the dataset). The base model is frozen, and only the FC layer is trained. The Macro F1 score is 0.69 for cardiffnlp/twitter-roberta-large-emotion-latest with this approach.

We also experimented with the entailment approach, in which the dataset was converted to a premise-hypothesis pair dataset with a label '0' or '1'. '0' for hypothesis being in contradiction/neutrality of the premise and '1' for hypothesis being entailment of the premise, respectively, for every emotion.

Example:

Input Sample: "But not very happy."

	Anger	**Fear**	**Joy**	**Sadness**	**Surprise**
Emotions	0	0	1	1	0

Converted to five different samples:

Premise	But not very happy.
Hypothesis	The speaker is feeling Anger.
Labels	[0] (Neutral or Contradiction)
Premise	But not very happy.
Hypothesis	The speaker is feeling Fear.
Labels	[0] (Neutral or Contradiction)
Premise	But not very happy.
Hypothesis	The speaker is feeling Joy.
Labels	[1] (Entailment)
Premise	But not very happy.
Hypothesis	The speaker is feeling Sadness.
Labels	[1] (Entailment)
Premise	But not very happy.
Hypothesis	The speaker is feeling Surprise.
Labels	[0] (Neutral or Contradiction)

Model Name	Track	Accuracy	Micro F1	Macro F1
SamLowe/roberta-base-go_emotions	A	0.21	0.45	0.44
cardiffnlp/twitter-roberta-large-emotion-latest	A	0.29	0.54	0.53
Emanuel/twitter-emotion-deberta-v3-base	A	0.16	0.45	0.40
meta-llama/Meta-Llama-3-8B-Instruct	A	0.24	0.58	0.59
cardiffnlp/twitter-roberta-large-emotion-latest (Full Fine-tuning)	A	0.43	0.60	0.49
SamLowe/roberta-base-go_emotions (Full Fine-tuning)	A	0.44	0.61	0.49
cardiffnlp/twitter-roberta-large-emotion-latest (Added Classifier Layer Only)	A	0.57	0.73	0.69
cardiffnlp/twitter-roberta-large-emotion-latest (Entailment Approach)	A	0.60	0.73	0.73
Fully fine-tuned Ensemble (final system evaluated on test set)	A	-	**0.7636**	**0.7344**

Table 1: Results of all the experiments conducted for Track A on Dev Set

The Premise and Hypothesis are then appended and given to a laguange model with an added final layer of single neuron to predict 0 or 1 for Contradiction and Entailment respectively.

All results are presented in Section-6.

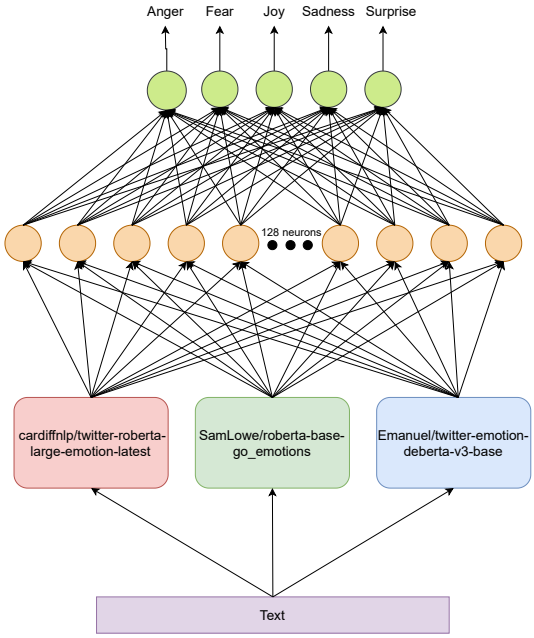

Figure 3: Model architecture of our system

5.2 Final Approach: Ensemble of Three Transformer Models

An ensemble using three transformer models: cardiffnlp/twitter-roberta-large-emotion-latest, SamLowe/roberta-base-go_emotions, and Emanuel/twitter-emotion-deberta-v3-base is created as shown in Figure-3. First, we tokenized the training, validation, and test datasets using the tokenizer for each model. A custom ensemble dataset was created to combine the inputs from all three models. While training, output of each model (logits) are concatenated together. The concatenated outputs were passed through two fully connected layers of 128 and 5 neurons for multi-label classification. Binary cross-entropy loss is used for training the model. The ensemble was trained for five epochs using a batch size of 8.

The training is done on an A30 24 GB GPU card. AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01 is employed. A linear learning rate scheduler with no warm-up steps is used for training. The models are trained with a maximum sequence length of 256 tokens. For regularization, dropout is set to 0.1 in the fully connected layers. The total training time for five epochs is approximately 30 minutes.

The results are analyzed in Section-6.

6 Results

The results of our experiments are summarized in Table 1. We began by evaluating several transformer-based pre-trained models without fine-tuning. We shortlisted models listed in for initial experiments and found that the macro-F1 score hovered between 0.40 to 0.50. To get a comparative understanding, we also evaluated a big model, meta-llama/Meta-Llama-3-8B-Instruct model, but could not achieve significant gains in the scores.

Afterward, we focused on smaller models and fine-tuned the cardiffnlp/twitter-roberta-

large-emotion-latest and SamLowe/roberta-base-go_emotions model. A fully fine-tuned version achieved a marginal improvement over the baseline. We also experimented with partial fine-tuning by adding only a classifier layer on top of twitter-roberta-large-emotion-latest, which resulted in the best performance among individual models and a 0.69 mmacro-F1 score. Employing an entailment-based approach with the cardiffnlp/twitter-roberta-large-emotion-latest model resulted in notable performance improvements, achieving a macro-F1 score of 0.73.

In our final approach, to further boost performance, we created an ensemble of pre-trained models trained on different emotion datasets and fully fine-tuned the ensemble with added classifier layers at the top. The models used were cardiffnlp/twitter-roberta-large-emotion-latest, SamLowe/roberta-base-go_emotions, and Emanuel/twitter-emotion-deberta-v3-base. The ensemble achieved the best overall performance with a Macro-F1 of 0.7344 in the test dataset, demonstrating the advantage of leveraging multiple models for emotion classification.

7 Conclusion

In this challenge, we explored various transformer-based models for multi-label emotion detection, evaluating their performance on Track A using multiple fine-tuning strategies and model ensembles. Our results demonstrate that model architecture and fine-tuning approach significantly impact performance. Our findings highlight the advantages of leveraging multiple pre-trained models and ensembling techniques for emotion classification tasks. Future work could explore additional architectures, data augmentation methods, and domain adaptation techniques to further enhance model performance and generalizability across different datasets. Our findings show that fine-tuning smaller models can sometimes perform as well as, or even better than, larger models. This means it is possible to improve accuracy without the requirement of very big models like LLMs.

Future Work

In the future, there are several ways we can improve the performance of our model. One important area is making these methods work for low-resource languages, where pre-trained models are not available. This might require new techniques like transfer learning or data augmentation techniques. We propose investigating cross-lingual transfer learning, e.g., XLM-R (Conneau et al., 2020), mBERT (Devlin et al., 2019) adaptations and back-translation-based data augmentation (Sennrich et al., 2016) to synthetically expand training data. Another challenge we face is class imbalance, where some emotions are harder to detect because they appear less often in the data. Future research could look at better ways to deal with this, such as creating more data for rare emotions e.g., SMOTE, Chawla et al. (2002) or using different loss functions, such as focal loss, Lin et al. (2018). Lastly, combining text with other types of data, like speech or images, as in multimodal fusion (Baltrušaitis et al., 2019) could make the model even better at understanding and classifying emotions. This could help create a more complete system for emotion detection.

Limitations

One of the key limitations of our approach is the lack of extensive focus on data augmentation strategies, which could have further enhanced model performance. While we explored an entailment-based reformulation to expand the dataset, we did not experiment with other augmentation techniques such as back-translation, synonym replacement, or adversarial data augmentation, which might have introduced greater diversity in training examples. The high computational cost of full fine-tuning is another constraint, as large transformer models require significant GPU resources, making scalability an issue. Better optimization techniques can mitigate the issue to a significant level.

Acknowledgments

This work was carried out at the Exploration Lab, IITK. We are thankful to the organisers of the competition for their contribution to the research community.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Frank A Acheampong, Henry Nunoo-Mensah, and Wei Chen. 2020. Transformer models for text-based

emotion detection: A comparative analysis. *arXiv preprint arXiv:2004.13704*.

Asim Ameer, Sher Maqbool, and Ghulam Azam. 2021. Multi-label emotion classification on code-mixed roman urdu and english sms messages using machine learning and deep learning approaches. In *2021 International Conference on Computer and Communication Technologies (ICCCT)*, pages 80–86. IEEE.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Ankush Chatterjee, Khyathi Raghavi Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

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.

Alan S Cowen and Dacher Keltner. 2018. Vocal expression of emotion reveals cross-cultural recognition, stereotypes, and differentiation. *Nature Human Behaviour*, 2(6):360–372.

Alan S Cowen and Dacher Keltner. 2019. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909.

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 (NAACL)*.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

X. Feng, K. Hui, and X. Deng et al. 2021. Understanding how the semantic features of contents influence the diffusion of government microblogs: Moderating role of content topics. *Information Management*, 58(8):103547.

A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183, Sofia, Bulgaria.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

A. Hu and S. Flaxman. 2018. Multimodal sentiment analysis to explore the structure of emotions. In *KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 350–358.

Emanuel Huber. 2021. Emanuel/twitter-emotion-deberta-v3-base.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

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*.

Sam Lowe. 2022. Samlowe/roberta-base-go_emotions.

Saif 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.

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.

Michele Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2020. Emotion recognition through a multi-model approach: A study of different word embedding techniques for emotion detection in textual data. In *Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 45–51. IEEE.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

M. Tasmin. 2018. Multi-dimensional aspect analysis of text input through human emotion and social factors. In *UbiComp '18: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 1779–1781.

Asalah Thiab, Luay Alawneh, and Mohammad AL-Smadi. 2024. Contextual emotion detection using ensemble deep learning. *Computer Speech Language*, 86:101604.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.