QiMP at SemEval-2025 Task 11: Optimizing Text-based Emotion Classification in English Beyond Traditional Methods

Mariia Bogatyreva, Pascal Gaertner, Quim Ribas Martinez, Daryna Dementieva, and Alexander Fraser
Technical University of Munich
mariia.bogatyreva@tum.de, pascal.gaertner@tum.de, quim.ribas01@estudiant.upf.edu,
daryna.dementieva@tum.de, alexander.fraser@tum.de

Abstract

As human-machine interactions become increasingly natural through text, accurate emotion recognition is essential. Detecting emotions provides valuable insights across various applications. In this paper, we present our approach for SemEval-2025 Task 11, Track A, which focuses on multi-label text-based detection of perceived emotions. Our system was designed for and tested on English language text. To classify emotions present in text snippets, we initially experimented with traditional techniques such as Logistic Regression, Gradient Boosting, and SVM. We then explored state-of-the-art LLMs (OpenAI o1 and DeepSeek V3) before developing our final system, utilizing a fine-tuned Transformer-based model. Our best-performing approach employs an ensemble of fine-tuned DeBERTa-large instances with multiple seeds, optimized using Optuna and StratifiedKFold cross-validation. This approach achieves an F1-score of 0.75, demonstrating promising results with room for further improvement. Additionally, this paper provides benchmark for 30 emotion classification methods on the BRIGHTER-English dataset.

1 Introduction

SemEval-2025 Task 11 (Muhammad et al., 2025b) focuses on detecting perceived emotion in short text snippets. Emotion detection has valuable applications in fields such as healthcare, education, finance (Hajek and Munk, 2023), customer service, and other applied domains (Kusal et al.; Liu et al.). Our participation in this task was centered on the English language, where we explored various approaches before ultimately adopting DeBERTa-large (He et al., 2021).

This task provides insights into how both humans and machines perceive emotions in written text, particularly in context-free scenarios. Our focus is on Track A, which involves identifying the presence of emotions in sentences without their intensity.

As multimodal research continues to advance, integrating text, visual, and audio modalities, text remains a critical component of emotion-recognition systems (Cheng et al.). Improving emotion detection in text-based snippets not only enhances classification accuracy but also better informs the selection of key triggers for emotional shifts.

Furthermore, text-based communication remains the dominant form of online interaction.

We approach this problem with a plethora of natural language processing techniques, from traditional methods to modern Large Language Models (LLMs) such as DeepSeek V3 (DeepSeek-AI et al., 2025) and OpenAI o1 (Jaech et al., 2024). The dataset for English language for this task is heavily imbalanced, making classification challenging. Performance is evaluated using the F1-score, which allows for a balanced identification of emotion's presence and absence.

We define *traditional* methods as those relying on rule-based, feature-based, lexicon-based, or static-embedding classifiers, such as models using TF-IDF, bag-of-words, or pre-trained word embeddings like GloVe, combined with algorithms such as SVM, Logistic Regression, MLP, or Gradient Boosting. *Beyond traditional* refers to transformer-based, pre-trained, or prompt-based models-such as BERT, DeBERTa, OpenAI o1, and DeepSeek V3, which leverage deep contextual representations, large-scale transfer learning, and, in the case of models like OpenAI o1 and DeepSeek V3, zero-shot inference capabilities. (Garrido-Merchan et al., 2023; Zhao et al., 2022)

To streamline experimentation, we initially used BERT-base (Devlin et al., 2018) before applying our optimizations to our best-performing model, DeBERTa-large. This allowed us to iterate efficiently while ensuring improvements translated effectively to our final model.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2283–2296

Text	Anger	Fear	Joy	Sadness	Surprise
But not very happy.	0	0	1	1	0
Well she's not gon na last the whole song like that, so since I'm behind her and the audience can't see below my torso pretty much, I use my hand to push down on the lid and support her weight.	0	0	1	0	0
She sat at her Papa's recliner sofa only to move next to me and start clinging to my arms.	0	0	0	0	0
Yes, the Oklahoma city bombing.	1	1	0	1	1
They were dancing to Bolero.	0	0	1	0	0

Table 1: Emotion annotations for text samples, 0 is for absence of emotion and 1 is for its presence

Throughout this task, we encountered several counterintuitive challenges. These challenges are discussed in Section 3, along with our hypotheses regarding their causes.

Our final system achieved an F1-score of 0.7537, placing us 27th out of 96 participating teams in the English track.

An analysis of our model's performance reveals notable class detection disparities. Anger, the least common emotion in our dataset, suffers from under-detection with the lowest recall (61.04%), meaning 38.96% of anger instances were missed. Conversely, fear, the most prevalent emotion, exhibits over-prediction tendencies, achieving excellent recall (91.20%) but the lowest specificity (64.77%). Despite these challenges, our model demonstrates strong overall performance, achieving multi-label subset accuracy of 47.04%, meaning the exact combination of emotions was correctly predicted in nearly half of all cases.

2 Background

2.1 Task 11 Bridging the Gap in Text-Based Emotion Detection (Track A Multi-label Emotion Detection)

In Task 11, Track A, we focused on detecting whether the emotion is present in a given sentence in English. This task revolves around perceived emotions, the emotions that most people are likely to infer from a short snippet of text provided without any context.

2.2 Related Research

In recent years, there has been significant research in this area of NLP, with various approaches proposed. These range from traditional Machine Learning techniques, such as Logistic Regression, Neural Networks, and XGBoost to Transformer-based models (Vaswani et al., 2023) like BERT, RoBERTa, DeBERTa. More recently, LLMs such as GPT-3.5, LLama 3, Mistral 7B, and Zephyr have been explored for emotion detection.

Initially, we were curious to see how more traditional and lightweight approaches would perform on this dataset. We experimented with different preprocessing and tokenization techniques, including tf-idf, word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014). Additionally, we used the NRC Emotion Lexicon (Mohammad and Turney, 2013) and the Sentiment Intensity Analyzer from NLTK. As shown in Subsection 3.1, even when incorporating Transformer-based tokenizers, the results were underwhelming.

Further research indicated that Transformer-based models consistently delivered the best performance. This led us to experiment with BERT, RoBERTa, DeBERTa, and other Transformer-based models. For the final system, we used a homogenous ensemble, leveraging initialization variance rather than architectural diversity.

2.3 Task Setup

Track A focuses solely on the presence (or absence) of emotions in a sentence, without considering their intensity. The dataset for the English language track covers five emotions: anger, fear, joy, sadness, and surprise. Since this is a multi-label classification task, any given sentence can express multiple emotions simultaneously.

Our final system was trained on both the train-

ing and development datasets, totaling 2,884 sentences. As noted in the competition paper, the English language data was sourced from social media, primarily from Subreddits such as *r/IAmA*. The sentences were annotated by multiple annotators using *Mechanical Turk* (Muhammad et al., 2025a). No additional datasets were utilized for model training.

3 System Overview

3.1 Experiments

See Appendix E, Table 3 for a ranking of F1 scores and correlating precision and recall scores across all our experiments, including our final model.

Preparation. Our initial investigation focused on data exploration and visualization to understand the dataset's characteristics. We conducted various analyses, including examining emotion distribution, computing the correlation matrix between emotions, and analyzing text length distribution. These insights allowed us to observe the class imbalances and potential biases within the dataset. We found that **anger** was significantly underrepresented, whereas **fear** was the dominant emotion (see Appendix A, Figure 2).

Traditional ML. We first explored traditional machine learning methods for emotion classification. Using Word2Vec, tf-idf, and GloVe embeddings, we converted text into numerical representations and fed them into various classifiers, including Support Vector Machines (SVM), Neural Networks (MLP), Logistic Regression, and Gradient Boosting. Despite their computational efficiency, these models struggled to capture complex contextual relationships. While the Neural Network classifier showed moderate performance, this approach could not model complex contextual dependencies, underperforming compared to transformer-based models. This reinforced the necessity of using more advanced approaches.

Preprocessing. We experimented with different preprocessing techniques to assess their impact on classification performance: grammar recognition and correction through tagging and lemmatization, removal of stop words (e.g., *the*, *is*, *but*), and removal of non-alphabetic characters (punctuation, numbers, and special symbols). However, these modifications resulted in only negligible improvements in model performance.

Lexicon-based Approaches. Lexicon-based methods are widely used in sentiment analysis. We tested two approaches: NRC Emotion Lexicon, which has predefined mappings of words to specific emotions, and NLTK Sentiment Analyzer, a polarity-based sentiment classifier. Neither approach yielded significant improvements, likely due to the inability of predefined word mappings to capture nuanced emotional contexts in text.

Transformer-based Models. Given that transformer-based models have consistently outperformed traditional approaches in sentiment analysis, we evaluated pre-trained models such as BERT, DistilBERT (Sanh et al., 2019), DeBERTa, and RoBERTa, followed by fine-tuned versions of these models to assess the impact of domain adaptation. As expected, fine-tuned transformers outperformed their pre-trained counterparts. BERT became our baseline for further experimentation as a lightweight model with comparable results.

BERT-Tokenizer + Traditional Classifier. We also experimented with using a BERT tokenizer for text representation, followed by classification using various traditional techniques: Neural Networks (MLP), Logistic Regression, SVM, GradientBoosting (LightGBM), and k-Nearest Neighbours (KNN). The best-performing combinations involved a neural network or logistic regression classifier, but their macro F1-scores still lagged behind fine-tuned transformer models.

Addressing the Class Imbalance. Since our dataset exhibited significant class imbalance, we explored two mitigation strategies:

1. Data Augmentation. We applied synonym replacement and back-translation (Edunov et al., 2018) (via French) using OPUS-MT models (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020) to generate additional samples while preserving the linguistic patterns provided. However, this had minimal or even negative effects on performance. These methods often introduced semantic drift or unnatural phrasing, adding noise rather than reinforcing emotional cues. Emotional nuances, critical for multi-label classification, were easily distorted during augmentation.

For future work, more advanced techniques like conditional text generation, semi-supervised learning, or cost-sensitive training could offer better solutions by preserving emotion-specific contexts while addressing imbalance more effectively.

2. Loss function Adjustments. We experimented with Class-Weighted Binary Cross-Entropy (to penalize misclassification of underrepresented emotions) and Focal Loss. Neither approach outperformed our baseline models.

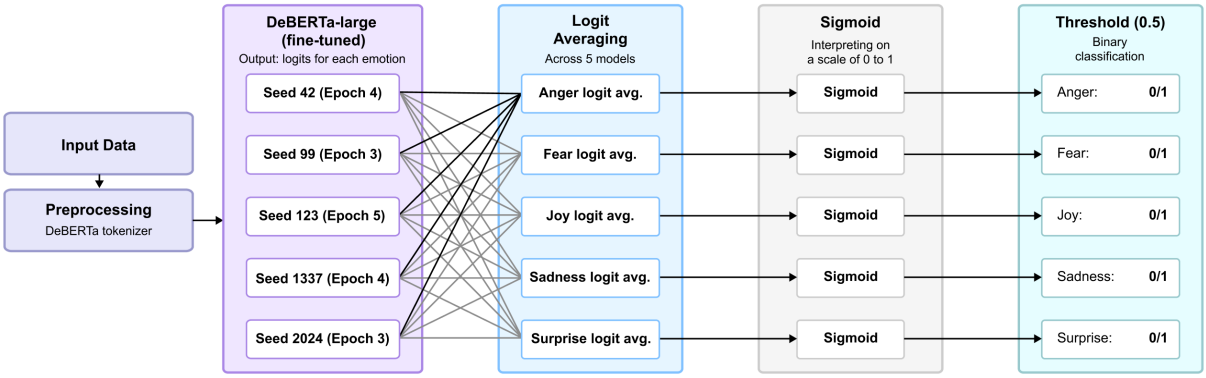

Figure 1: Runtime prediction pipeline for the final model

GPT-based LLMs. We chose to evaluate OpenAI o1 and DeepSeek V3 in a zero-shot setting, as it has been shown that few-shot prompting tends to perform even worse than zero-shot and fine-tuning (Kazakov et al.). These models performed worse than simpler methods like a BERT tokenizer with logistic regression. This is likely because zero-shot models rely on general pretraining rather than adopting the dataset's specific distribution and nuances. See Appendix D, Figure 14 for prompt details.

Ensembling. We explored ensembling by combining fine-tuned transformer models using various voting methods: SoftVoting, HardVoting, WeightedVoting, and Stacking (meta-learning). Ensembling had the most significant positive impact on predictive abilities, consistently outperforming individual transformer models.

Hyperparameter Optimization. Since ensembling yielded the best results, we further refined our approach using Optuna (Akiba et al., 2019), an open-source hyperparameter optimization framework. Optuna enabled an automated search for optimal configuration, tuning parameters such as learning rate, batch size, and dropout rate.

3.2 Final Model.

After extensive experimentation, DeBERTa-large consistently achieved the highest overall F1-scores, outperforming other transformer-based models like BERT, DistilBERT and RoBERTa.

To enhance robustness and reduce overfitting, we trained DeBERTa-large using five different seeds (42, 99, 123, 1337, and 2024). To maintain the class distribution of the original dataset, we applied stratified cross-validation with three folds.

Hyperparameter Selection. We used Optuna to optimize hyperparameters, resulting in the following configuration applied to all five seeds: *learning_rate* of 7.130877023256217e-06, *batch_size* of 8, *max_length* of 128, and *number_of_epochs* of 5.

Training and Prediction Process. Each model was trained for five epochs per seed. For the final ensemble prediction, we first computed the logits from each model (corresponding to different seeds), then averaged these logits across all models. Next, we applied the sigmoid function to obtain probability scores for each emotion. Finally, we assigned labels based on a threshold of 0.5: if the probability was ≥ 0.5, the emotion was considered present; otherwise, it was considered absent. Prediction pipeline can be seen in Figure 1.

Performance and Computational Cost. Fine-tuning DeBERTa-large improved generalization across underrepresented emotions, making it the most effective model for our final predictions. However, this approach was computationally expensive, requiring 435 million parameters. Due to our resource constraints, we were unable to test larger models.

We discuss performance-cost trade-off in-depth in Subsection 5.5.

4 Experimental Setup

4.1 Training Data

The dataset provided by the competition was split into train, dev, and test, consisting of 2,768, 116, and 2,767 sentences, respectively. For fine-tuning our model, we utilized both the train and dev datasets. The distribution of emotions across the combined dataset can be found in Appendix A, Figure 2.

4.2 Training Details

We fine-tuned the pre-trained DeBERTa-large model from HuggingFace. We trained five instances using different seeds (42, 99, 123, 1337, 2024) and trained them each for five epochs. The best epoch for each seed was 4, 3, 5, 4, and 3, respectively. Hyperparameters such as learning rate, batch size, and max length were optimized using Optuna. The stochastic optimization method used was AdamW.

4.3 Hardware and Hyperparameters

Our experiments were implemented using PyTorch 2.5.0, HuggingFace transformers 4.46.3 (for DeBERTa-large), and scikit-learn 1.5.1 (for evaluation and preprocessing). Model training was conducted using the free version of Google Colab, while the final model was trained fully on an Apple M2 Pro chip.

4.4 Evaluation Metrics

As specified by the shared task, we evaluated model performance using the macro-F1 score. All formulas mentioned throughout the paper can be found in Appendix D, Figure 13.

5 Results

5.1 Key Findings

Our final model demonstrated strong performance in multi-label emotion detection, particularly in identifying fear (91.20% recall) and joy (84.18% recall), however, it struggled with anger (61.04% recall), likely due to its underrepresentation in the dataset.

5.2 Main Quantitative Findings

Summary of recall (sensitivity) and specificity for each emotion can be found in table 5.2 summarizes.

Emotion	Recall (Sensitivity)	Specificity
Anger	0.6104	0.9704
Fear	0.9120	0.6477
Joy	0.8418	0.8759
Sadness	0.8079	0.8571
Surprise	0.7983	0.8540

Table 2: Performance of the model on the test set.

The class imbalance issue negatively impacted anger detection, as the low number of training samples made it harder for the model to learn meaningful patterns, resulting in lower recall.

To complement our macro F1 evaluation, further reports can be found in Appendix B, such as ROC-AUC (receiver operating characteristic, area under the curve) curves per emotion and Matthews Correlation Coefficient (MCC) per emotion, as well as precision-recall graphic. Our MCC scores range from 0.59 (fear) to 0.67 (joy), indicating strong and balanced performance across labels.

5.3 Error Data

Our confusion matrices and error analysis revealed several key insights:

- Fear was frequently over-predicted, leading to a high false positive rate (434 FP).
- Anger had the lowest recall (61.04%), likely due to its low representation in the dataset.
- False Positives: The model often overpredicted emotions, particularly in ambiguous snippets where multiple interpretations were possible.
- False Negatives: Implicit emotions (e.g., subtle anger) were often missed, suggesting that the model struggled with nuanced emotional expressions.

5.4 Error Analysis

An in-depth analysis of the misclassified sentences uncovered several patterns.

Ambiguous Phrasing. Many sentences had multiple valid emotional interpretations, making classification difficult. The model often assigned incorrect labels in these cases.

Sentence Length Variability. The dataset contained short, medium, and long sentences, adding complexity. Short sentences lacked emotional cues, while longer sentences often contained mixed emotions, both cases made classification harder.

Labeling Inconsistencies. Some annotations appeared counterintuitive, potentially introducing noise into the training process and reducing model accuracy.

Overprediction of Multiple Emotions. For single-label sentences, the model frequently predicted two emotions instead of one, indicating overlapping textual patterns. For multi-label sentences (three or more emotions), accuracy declined, with inconsistent predictions regarding the number of emotions present.

These findings suggest potential future improvements, including enhancing neutral sentence classification, refining multi-label prediction strategies,

and improving robustness to ambiguous text. Despite these challenges, our model demonstrated strong generalization and competitive performance, underscoring its effectiveness in detecting emotions in text.

5.5 Considerations for Balancing Accuracy and Efficiency

While our final system, based on an ensemble of fine-tuned DeBERTa-large models, achieved the best empirical performance, it introduced substantial computational costs. DeBERTa-large, with approximately 435 million parameters per model, combined with training multiple seeds, resulted in high memory and processing requirements.

Although we considered lighter alternatives such as DistilBERT during our model selection phase, we prioritized DeBERTa-large for its superior performance. Nevertheless, it is well-known that smaller models generally offer faster inference at the cost of reduced accuracy, presenting a trade-off between efficiency and performance.

Balancing performance and computational demands remains a critical challenge. Future work could explore approaches such as:

Model Compression Techniques. Add pruning or quantization in order to reduce model size without significant loss of accuracy.

Optimized Ensembling Strategies. Reducing the number of models combined or adopting techniques like snapshot ensembling to maintain robustness with lower resource usage.

Adopting these strategies could help retain strong predictive power while making the system more practical and scalable for real-world applications.

6 Conclusion

Our proposed approach, leveraging DeBERTa-large with multiple seeds, ensemble methods, Optuna hyperparameter optimization, and StratifiedKFold cross-validation, achieved an F1-score of 0.7537, surpassing the baseline provided by the task organizers. While our model demonstrated strong performance, our final ranking suggests room for improvement. This study explored traditional emotion recognition techniques, LLMs, and Transformer-based approaches, highlighting the successful application of advanced ensemble methods to Task 11 at SemEval-2025.

Future improvements may focus on exploring computationally efficient alternatives to DeBERTa-large for better scalability, expanding and balancing training data to reduce class imbalance issues, and implementing more robust labeling strategies, accounting for shortcomings of the current model discussed in Section 5.

While our research explored several widely used approaches, we recognize that other promising techniques could achieve comparable or better results, potentially with lower computational costs. These include lexicon-based approaches (NRC VAD (Garcia et al., 2024), SenticNet (Butt et al., 2021)), alternative transformer models (SiEBERT (Rozado et al., 2022)), LLM approaches (Zephyr (Shaik et al.), LLama 2, InstructERC (Cheng et al.; Lei et al., 2024)), more traditional machine learning methods (LSTM (Geethanjali and Valarmathi, 2024; Kumar et al.)), and explainability techniques like SHAP (Hajek and Munk, 2023; Butt et al., 2021).

Although our initial attempts at re-weighting and naïve augmentation did not yield significant results, future work could explore multi-label aware oversampling (e.g., MLSMOTE (Charte et al., 2015) or similarity-based oversampling (Karaman et al., 2024)), adaptive batch-level strategies, such as loss-driven batch selection (Loshchilov and Hutter, 2016; Zhou et al., 2024), and deferred re-weighting schedules (Cao et al., 2019), that apply stronger class weights only after an initial warm-up phase, to mitigate the severe class imbalance.

Moreover, addressing dataset biases remains crucial. Generalization is particularly challenging in English, where linguistic and cultural diversity influences both text production and emotional perception. Future work should extend beyond English to encompass multilingual and cross-cultural perspectives, incorporating sociolinguistic and anthropological insights. Additionally, integrating non-verbal elements like emojis into text-based emotion recognition may improve model robustness and real-world applicability.

7 Ethical Considerations

This study addresses perceived emotions rather than the true internal emotional states of users. Perception of emotion is inherently subjective, shaped by individual factors such as gender, culture, language, and personal experience. As such, we do not claim that our model's outputs reflect any universal or objective emotional truth.

Importantly, the data used in this task was sourced from online and social media contexts, which may introduce cultural, linguistic, and demographic biases. Such biases can affect both the generalization ability and fairness of the model, leading to underrepresentation or misinterpretation of emotions from diverse user groups. Additionally, the dataset was annotated by crowdsourced workers whose backgrounds are unknown, potentially reinforcing subjective or culturally specific patterns.

To promote more equitable development in emotion detection, we recommend to strive to capture a broader diversity of emotional expressions across different populations in the future, in order to mitigate risks of marginalization or misrepresentation.

Given these limitations, we strongly discourage the deployment of this system in high-stakes applications requiring precise emotional interpretation, such as clinical diagnostics, automated decision-making, or areas involving sensitive personal data. Potential misuses could include emotional manipulation, targeted social engineering, or exploitation of vulnerable groups, and developers should exercise caution in adapting the model to real-world settings.

8 Acknowledgment

This research was conducted as part of the Bachelor Lab Course on Natural Language Processing at TUM, under the guidance of Prof. Dr. Alexander Fraser's Data Analytics & Statistics Group at TUM School of Computation, Information and Technology. We extend our gratitude to the research group for providing us with the opportunity to participate in this competition and apply hands-on NLP techniques.

We are deeply grateful to the anonymous reviewers for their constructive feedback, which helped us improve the quality of this paper.

We also thank the SemEval-2025 Task 11 organizers for curating this challenge and advancing research in text-based emotion recognition. Additionally, we acknowledge the academic teams behind the BRIGHTER dataset, which bridges the gap in human-annotated emotion recognition resources by providing multi-label emotion data across 28 languages, thereby fostering innovation in multilingual emotion analysis.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *Preprint*, arXiv:1907.10902.

Sabur Butt, Shakshi Sharma, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. 2021. What goes on inside rumour and non-rumour tweets and their reactions: A Psycholinguistic Analyses. *arXiv preprint*. ArXiv:2112.03003 [cs].

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa-Anke, Fangyu Liu, Eugenio Martínez-Cámara, et al. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E. Association for Computational Linguistics.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Preprint*, arXiv:1906.07413.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397.

Zebang Cheng, Fuqiang Niu, Yuxiang Lin, Zhi-qi Cheng, Xiaojiang Peng, and Bowen Zhang. MIPS at SemEval-2024 task 3: Multimodal emotion-cause pair extraction in conversations with multimodal language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 667–674. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *Preprint*, arXiv:1808.09381.

Santiago Garcia, Elizabeth Martinez, Juan Cuadrado, Juan Martinez-santos, and Edwin Puertas. 2024. VerbaNexAI lab at SemEval-2024 task 10: Emotion recognition and reasoning in mixed-coded conversations based on an NRC VAD approach. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1332–1338. Association for Computational Linguistics.

Eduardo C. Garrido-Merchan, Roberto Gozalo-Brizuela, and Santiago Gonzalez-Carvajal. 2023. Comparing bert against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*, 2(4):352–356.

R. Geethanjali and A. Valarmathi. 2024. A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media. 14(1):22270.

Petr Hajek and Michal Munk. 2023. Speech emotion recognition and text sentiment analysis for financial distress prediction. 35(29):21463–21477.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

OpenAI: Aaron Jaech, Adam Kalai, et al. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Ismail Hakki Karaman, Gulser Koksal, Levent Eriskin, and Salih Salihoglu. 2024. A similarity-based oversampling method for multi-label imbalanced text data. *Preprint*, arXiv:2411.01013.

Roman Kazakov, Kseniia Petukhova, and Ekaterina Kochmar. PetKaz at SemEval-2024 task 3: Advancing emotion classification with an LLM for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1127–1134. Association for Computational Linguistics.

Taewoon Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *Preprint*, arXiv:2108.12009.

Shivani Kumar, Md. Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. SemEval 2024 - task 10: Emotion discovery and reasoning its flip in conversation (EDiReF). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1933–1946. Association for Computational Linguistics.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. 56(12):15129–15215.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *Preprint*, arXiv:2309.11911.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. Emotion classification for short texts: an improved multi-label method. 10(1):306.

Ilya Loshchilov and Frank Hutter. 2016. Online batch selection for faster training of neural networks. *Preprint*, arXiv:1511.06343.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Preprint*, arXiv:1310.4546.

Saif M. Mohammad and Peter D. Turney. 2013. Nrc emotion lexicon. Technical report. 234 p.

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 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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

David Rozado, Ruth Hughes, and Jamin Halberstadt. 2022. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. 17(10):e0276367.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version

of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Zuhair Hasan Shaik, Dhivya Prasanna, Enduri Jahnavi, Rishi Thippireddy, Vamsi Madhav, Sunil Saumya, and Shankar Biradar. FeedForward at SemEval-2024 task 10: Trigger and sentext-height enriched emotion analysis in multi-party conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 745–756. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Liping Zhao, Waad Alhoshan, Alessio Ferrari, and Keletso J. Letsholo. 2022. Classification of natural language processing techniques for requirements engineering. *Preprint*, arXiv:2204.04282.

Ao Zhou, Bin Liu, Jin Wang, and Grigorios Tsoumakas. 2024. Multi-label adaptive batch selection by highlighting hard and imbalanced samples. *Preprint*, arXiv:2403.18192.

A Appendix

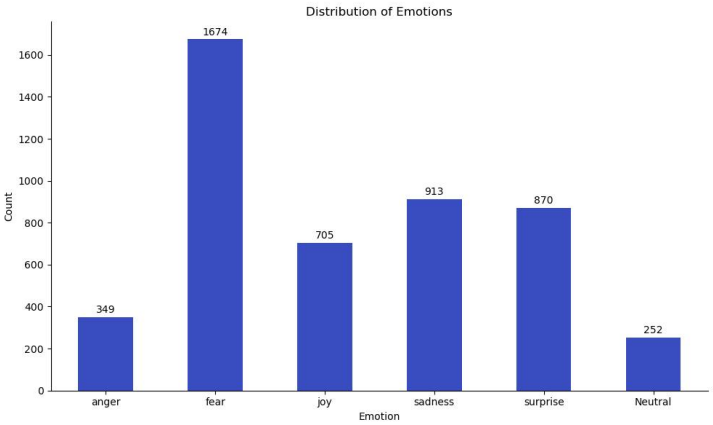

Figure 2: Emotion Distribution in training and development data

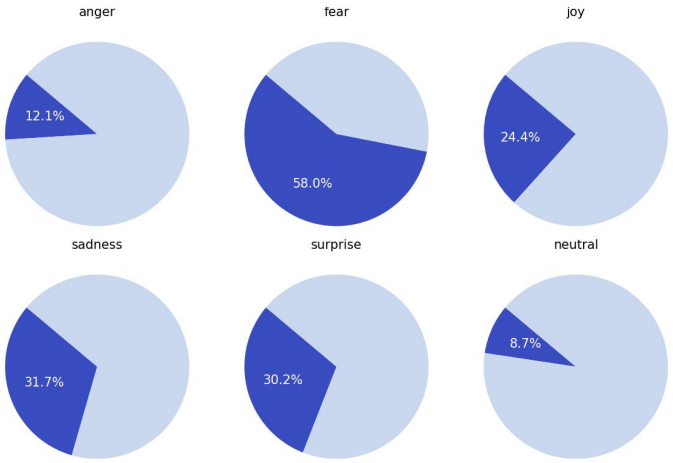

Figure 3: Emotion Distribution

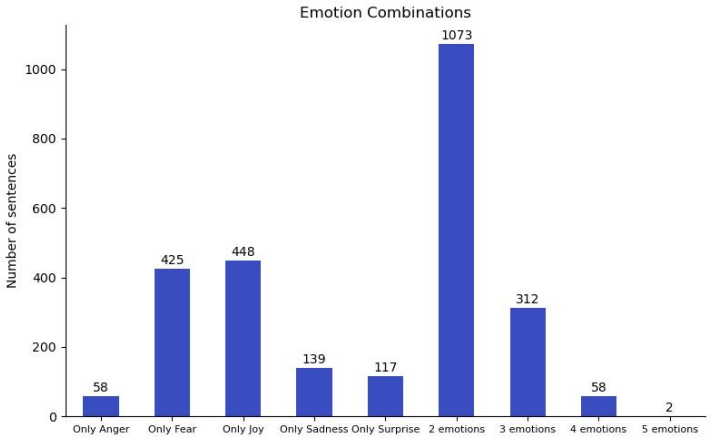

Figure 4: Label Sentence Distribution

B Appendix

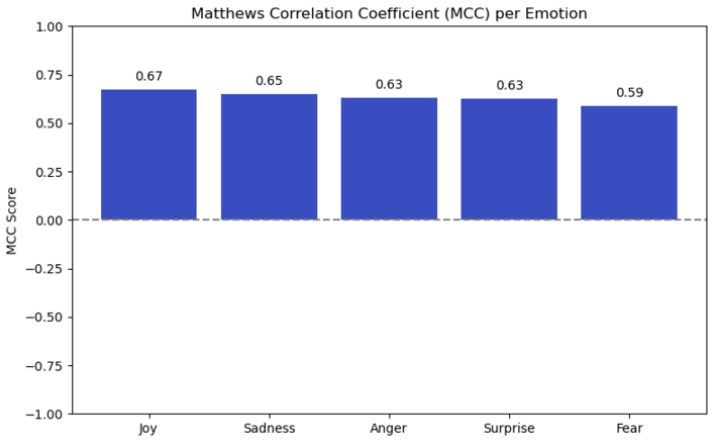

Figure 5: Matthews Correlation Coefficient (MCC) per Emotion

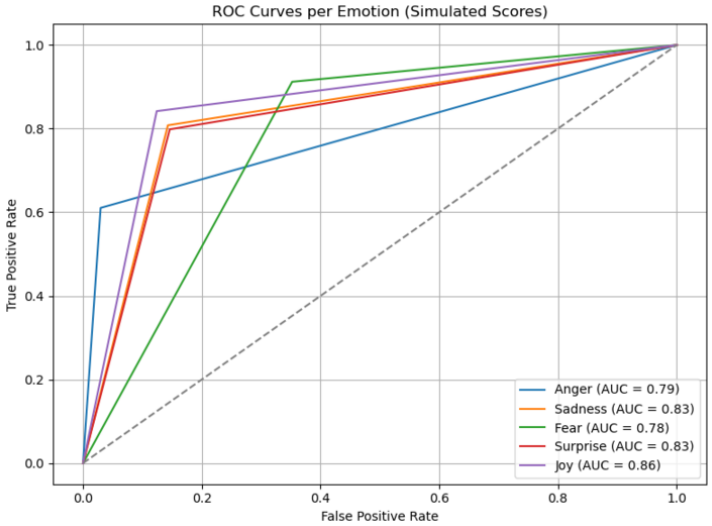

Figure 6: ROC-AUC Curves per Emotion

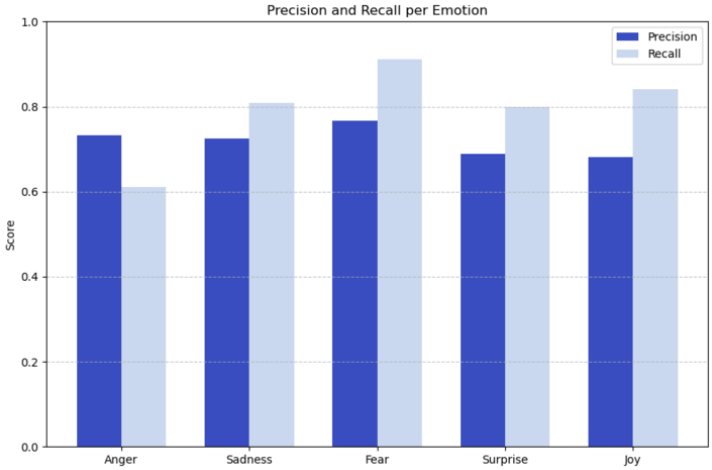

Figure 7: Precision and Recall per Emotion

C Appendix

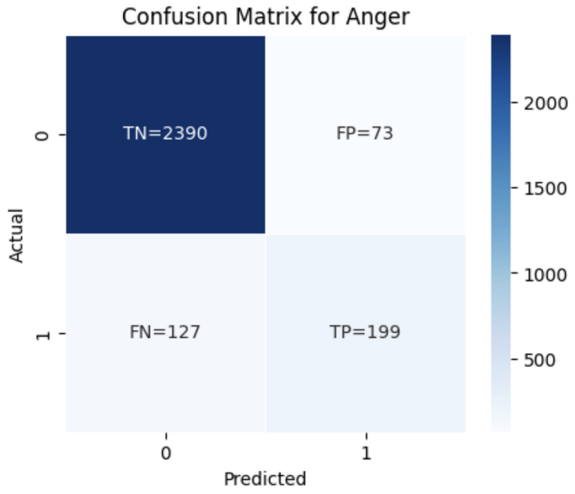

Figure 8: Confusion Matrix for Anger

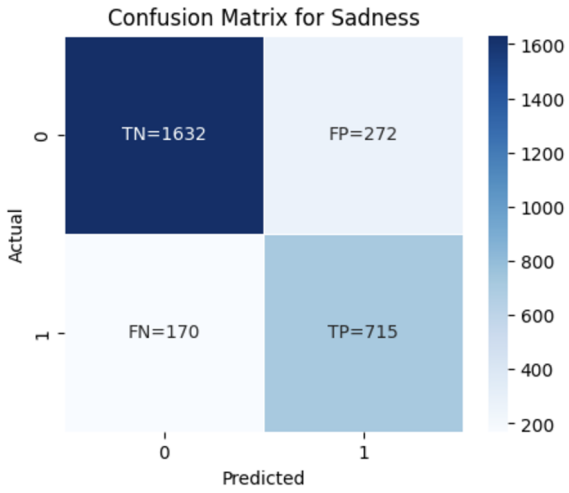

Figure 11: Confusion Matrix for Sadness

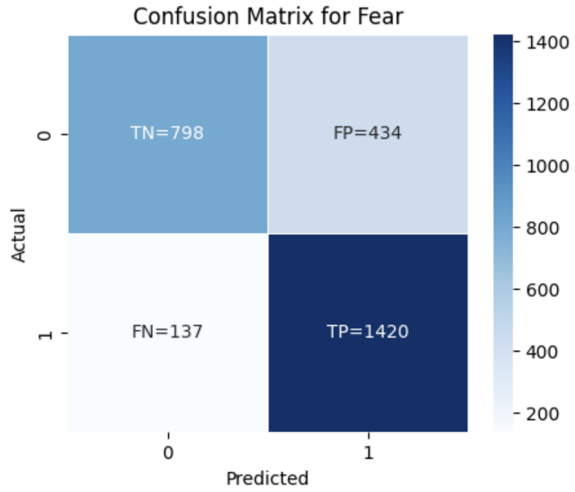

Figure 9: Confusion Matrix for Fear

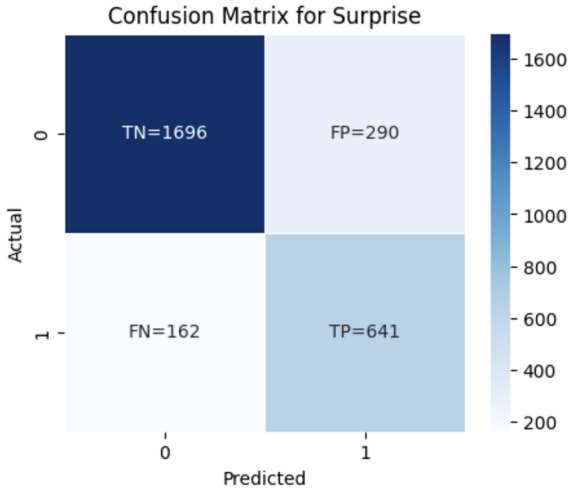

Figure 12: Confusion Matrix for Surprise

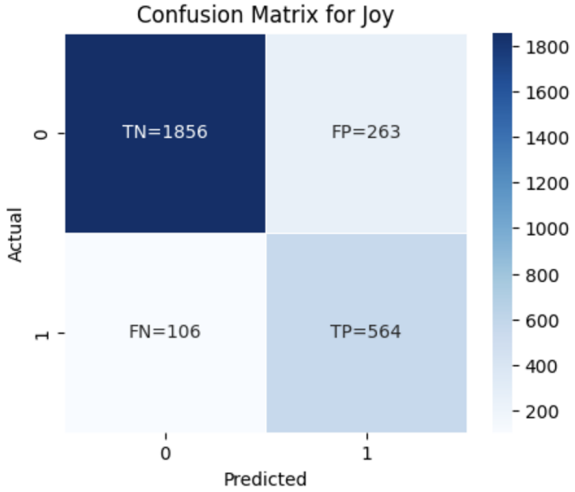

Figure 10: Confusion Matrix for Joy

D Appendix

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \qquad \text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \qquad F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \qquad \text{Macro } F1\text{-Score} = \frac{\sum_{i=1}^{n} F1\text{-Score}_i}{n}$$

Figure 13: Formula definitions for metrics used in our evaluation. High precision indicates a low rate of false positives. High recall means the model identified most positive samples correctly. Specificity measures the model's ability to correctly identify negative samples. F1 is useful when balancing precision and recall, especially with imbalanced classes. Accuracy represents the overall correctness across all predictions. The macro F1 Score is the average of F1 scores for each class.

prompt: Now you are an expert on sentiment and emotional analysis. Please infer, considering the content and the way the sentence is written, what emotions from a list of [anger, fear, joy, sadness, surprise], if any, are perceived from this sentence by the majority of people
format: Return your answer in .csv format "sentence, 0/1, 0/1, 0/1, 0/1, 0/1", where 0 represents absence of emotion and 1 its presence
warning: Don't hallucinate and find the emotions the majority of people would agree with
target: [sentence S_i].

Figure 14: Zero-shot prompt template for LLMs

E Appendix

Rank	Model (# of epochs if finetuned)	Precision	Recall	F1-score
1	DeBERTa-large (5) + Multiple seeds + Ensemble + Optuna + StratifiedKFold	0.81	0.75	0.76
2	RoBERTa+DeBERTa+BERTweet (5) + Ensemble	0.75	0.73	0.74
3	DeBERTa (3)+ Multiple seeds + Ensemble + Optuna + StratifiedKFold	0.75	0.71	0.73
4	DeBERTa (5)	0.70	0.75	0.72
5	DeBERTa (5) + Gridsearch hyperparameters + Weight based oversampling	0.70	0.73	0.71
6	cardiffnlp/twitter-roberta-base-emotion-multilabel-latest (3) (Camacho-Collados et al., 2022)	0.69	0.72	0.70
7	BERT (4)	0.71	0.70	0.70
8	RoBERTa (5)	0.72	0.68	0.70
9	cardiffnlp/twitter-roberta-base-emotion-multilabel-latest (3) + Contrastive loss	0.69	0.69	0.69
10	BERTweet (4)	0.76	0.65	0.69
11	RoBERTa (3) + Synonym data augmentation	0.75	0.64	0.69
12	BERT (2) + Back translation data augmentation	0.75	0.63	0.68
13	EmoBERTa (5) (Kim and Vossen, 2021)	0.72	0.65	0.68
14	BERT + BCE (4)	0.77	0.57	0.65
15	BERT	0.78	0.59	0.65
16	DistilBERT	0.73	0.60	0.65
17	BERT Tokenizer + Neural Networks (MLP)	0.66	0.59	0.62
18	BERT Tokenizer + Logistic Regression	0.65	0.58	0.61
19	BERT Tokenizer + SVM	0.57	0.58	0.58
20	OpenAI o1 zero-shot	0.97	0.37	0.58
21	DeepSeek V3 zero-shot	1.00	0.39	0.54
22	DistilBERT (3) + contrastive loss	0.70	0.46	0.49
23	BERT Tokenizer + Gradient Boosting (LightGBM)	0.74	0.42	0.48
24	BERT (1) + Focal loss	0.31	1.00	0.45
25	BERT Tokenizer + KNN	0.56	0.41	0.44
26	SentimentIntensityAnalyzer from NLTK + Preprocessing	0.47	0.44	0.43
27	GloVe 6B 100d + Logistic Regression	0.63	0.37	0.43
28	NRC + Preprocessing	0.48	0.44	0.43
29	tf-idf + Gradient Boosting (XGBoost)	0.56	0.34	0.38
30	word2vec + Gradient Boosting (XGBoost)	0.37	0.25	0.27

Table 3: Performance comparison of all models based on macro F1-score. Number of epochs presented is the best for specific models