MRS at SemEval-2025 Task 11: A Hybrid Approach for Bridging the Gap in Text-Based Emotion Detection

Milad Afshari **Richard Frost** **Samantha Kissel** **Kristen Marie Johnson**
Michigan State University
{afsharim, frostric, kisselsa, kristenj}@msu.edu

Abstract

We tackle the challenge of multi-label emotion detection in short texts, focusing on SemEval-2025 Task 11 Track A. Our approach, *RoEmo*, combines generative and discriminative models in an ensemble strategy to classify texts into five emotions: anger, fear, joy, sadness, and surprise. The generative model, instruction-finetuned on emotion detection datasets, undergoes additional fine-tuning on the SemEval-2025 Task 11 Track A dataset to enhance its performance for this specific task. Meanwhile, the discriminative model, based on binary classification, offers a straightforward yet effective approach to classification. We review recent advancements in multi-label emotion detection and analyze the task dataset. Our results show that RoEmo ranks among the top-performing systems, demonstrating high accuracy and reliability.

1 Introduction

Emotion detection plays a critical role in natural language processing (NLP), yet remains a challenging task due to the nuanced and often subjective nature of emotional content. While recent advancements in emotion detection have demonstrated significant progress, there is still a need to enhance both the accuracy and efficiency of these models (Seyeditabari et al., 2018).

Emotion detection in short text settings presents a significant challenge due to the limited contextual information available, which constrains traditional NLP models that typically rely on richer textual input (Pang et al., 2021). The scarcity of extensive context necessitates the development of specialized approaches capable of accurately capturing emotional content from minimal linguistic cues. This challenge becomes even more pronounced in multi-label emotion detection, where the task involves identifying the emotion that most people perceive in the speaker's words—rather than the reader's reaction, the emotion of other characters, or the speaker's true feelings. The inherent subjectivity of language interpretation and the ambiguity in emotional expression often lead to multiple valid interpretations, making the task even more complex (Figure 1).

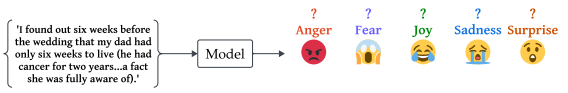

Figure 1: The challenge of multi-label emotion detection. This task focuses on identifying the emotion that most people perceive in the speaker's words, rather than the reader's reaction, the emotion of other characters, or the speaker's true feelings. The difficulty arises from the subjective nature of language interpretation and the inherent ambiguity in emotional expression.

Our study centers on SemEval-2025 Task 11 (Muhammad et al., 2025b), a benchmark challenge in emotion detection for short texts. Specifically, we address Track A, which requires classifying texts into one or more of five emotions: anger, fear, joy, sadness, and surprise. While Track A includes multiple languages, our focus is exclusively on English. Our approach, RoEmo, combines the predictions of RoBERTa-large (Liu et al., 2019) and EmoT5 (Liu et al., 2024) to enhance emotion detection in short texts.

2 Related Works

Emotion detection in text is a long-studied task in natural language processing with several approaches achieving success over the years. In this section, we summarize the most common emotion models employed in emotion detection and recent successful strategies for emotion detection.

2.1 Emotion Models

There are several competing models for emotions commonly used in emotion detection tasks, each

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 584–589

offering a unique perspective on how to conceptualize and classify affective states. For instance, the SemEval-2025 dataset used for this task is built upon the model proposed by Ekman (1992), which posits the existence of six basic emotions: joy, sadness, fear, anger, surprise, and disgust. Another well-known model proposed by Plutchik (2001) expands to eight primary emotions, introducing a broader perspective on affective states. While these models emphasize identifying base emotional states, some researchers have opted to conceptualize emotions within a multi-dimensional space. A notable example is the approach by Russell and Mehrabian (1977), which models emotions as three-dimensional vectors characterized by valence, arousal, and dominance.

Together, these models illustrate the diversity in theoretical approaches to understanding emotions. The choice between these frameworks often depends on the specific objectives and constraints of the emotion detection task at hand, with each model bringing its own strengths and limitations to the field.

2.2 Emotion Detection Methodologies

There have been several approaches that have found success in emotion detection. A notable characteristic of many approaches to emotion detection is the use of emotional lexicons to create additional model features (Al Maruf et al., 2024). These lexicons are simply lists of words associated with a particular emotion. A popular example is the EmoLex lexicon due to Mohammad and Turney (2010), which contains over 2000 terms along with their (Plutchik, 2001) emotion associations.

The top performer in SemEval-2018's task 1, which evaluated emotion detection in tweets, extracted five feature vectors from each tweet which were then processed by several models before being combined using ensemble methods (Duppada et al., 2018). More recently, Huang et al. (2021) considered a multi-label emotion classification problem similar to our own using a bi-directional LSTM encoder-decoder model. This approach was inspired by earlier works such as (Godbole and Sarawagi, 2004) and (Read et al., 2011), which advocated for the use of series of simple binary classifiers for multi-label classification problems. Other notable approaches have focused on exploiting nontraditional emotion models (Casel et al., 2021) and transfer learning based approaches (Yu et al., 2018). Graphical Neural Networks, such as EmoGraph (Xu et al., 2020), and span-prediction approaches, including SpanEmo (Alhuzali and Ananiadou, 2021), represent additional advancements in the field.

Recently, the EmoLLM framework (Liu et al., 2024) has leveraged instruction-tuned large language models (LLMs), such as EmoLLaMA, for emotion detection, demonstrating enhanced performance in both categorical and regression-based affective tasks. The EmoLLM series marks a substantial advancement in emotion detection by utilizing large language models fine-tuned on a multi-task affective instruction dataset. This approach enables EmoLLMs to excel in complex emotion analysis tasks, achieving performance levels comparable to, or better than, leading models such as ChatGPT and GPT-4 (Liu et al., 2024).

3 Approach

In this work, we employ two transformer-based models, **RoBERTa-large** and **EmoT5**, for emotion detection. Our approach explores both discriminative and generative modeling paradigms to classify emotions effectively. Both models are finetuned for 15 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} on an NVIDIA RTX A6000 GPU.

3.1 RoBERTa-large with MLP

RoBERTa-large is a transformer model optimized for masked language modeling, providing robust contextual representations. To adapt it for emotion classification, we append a *two-layer multi-layer perceptron (MLP)* on top of the final hidden states. This additional MLP enables the model to refine its learned features for classification.

We formulate emotion detection as a **binary classification problem** for each emotion. Given an input text, the model predicts the probability of each emotion being present, allowing for multi-label classification. The MLP layers are trained jointly with RoBERTa using a binary cross-entropy loss function.

3.2 EmoLLM-based Approach

Recent advancements in LLMs have significantly improved affective computing tasks, particularly in emotion detection. One notable development in this domain is **EmoLLMs**, a class of *instruction-tuned* LLMs specifically fine-tuned for emotion recognition (Liu et al., 2024). These models uti-

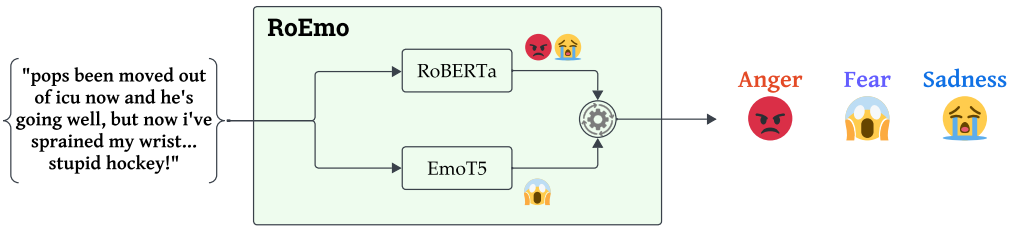

Figure 2: **Overview of RoEmo**: We obtain predictions from the discriminative model RoBERTa-large and the generative, instruction-finetuned model EmoT5. The results from these two models are then combined using a logical OR operation, leveraging their predictions to determine each emotion.

lize the first multi-task Affective Analysis Instruction Dataset (AAID), which includes the SemEval-2018 Task 1: Affect in Tweets dataset (Mohammad et al., 2018) along with other datasets. Additionally, they leverage an Affective Evaluation Benchmark (AEB) designed to measure affective generalization. Using these resources, the authors developed instruction-following LLMs optimized for diverse affective analysis tasks.

Empirical evaluations show that EmoLLMs achieve state-of-the-art (SOTA) performance on AEB, outperforming other open-source models. Additionally, they exhibit generalization capabilities on par with the GPT family of models, establishing their potential as highly effective tools for emotion detection. Their ability to process complex affective cues makes them strong candidates for real-world applications requiring fine-grained emotion classification (Liu et al., 2024).

Among the models evaluated in the EmoLLMs framework, we selected *EmoT5* for its superior performance in emotion classification. EmoT5 follows a *generative* approach, framing the task as text-to-text generation, where it directly generates emotion labels based on the input text.

Following (Liu et al., 2024), we structured the classification process using the following prompt template:

```
Task: [task prompt] Tweet: [input text] This tweet contains emotions: [output]
```

With the task prompt:

> Categorize the tweet's emotional tone as either 'neutral or no emotion' or identify the presence of one or more of the given emotions (anger, fear, joy, sadness, surprise).

Fine-tuning EmoT5 with this setup optimized its performance for our emotion detection task.

3.3 RoEmo: A Hybrid Approach

While both RoBERTa-large and EmoT5 individually offer strong performance in emotion detection, we observe that their predictions often differ, highlighting their complementary strengths. To leverage the strengths of both models, we propose RoEmo, a hybrid approach that combines their predictions. Specifically, we apply a **logical OR** operation to merge the outputs, predicting an emotion as present if either model detects it (see Figure 2). This simple yet effective fusion strategy allows the model to capture a broader range of emotional cues, improving its ability to detect emotions that might be overlooked by one of the models alone.

Empirical results show that RoEmo performs especially well when emotions are subtle or ambiguous. By combining predictions from RoBERTa and EmoT5, the ensemble increases *recall*, as it is less likely to miss true positive cases. Although this may slightly reduce *precision*—since more predictions can introduce some noise—the overall *macro-F1 score* improves. This highlights the complementary strengths of the two models and demonstrates the effectiveness of hybrid modeling in emotion classification tasks.

4 Experiments

In this section, we analyze the dataset to understand its characteristics and challenges. We then present the performance of RoBERTa-large, followed by EmoT5, and finally, evaluate the combined output using our ensemble method, RoEmo. We highlight the benefits of this hybrid approach in improving multi-label emotion detection.

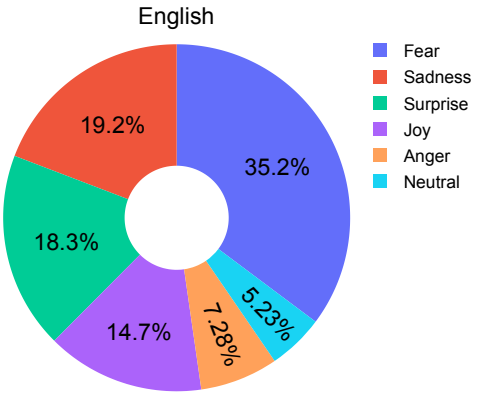

Figure 3: The distribution of emotion labels in the training data. Note that some emotions can co-occur—for example, a single text may be labeled with both joy and surprise.

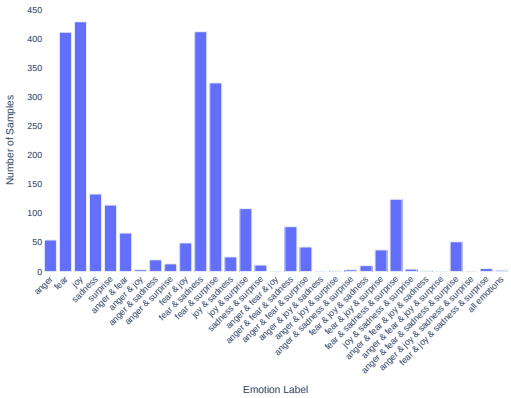

Figure 4: Bar plot depicting the distribution of samples according to the emotion labels assigned to each.

4.1 Dataset overview

The dataset for SemEval-2025 Task 11 (Track A) is designed for multi-label emotion detection (Muhammad et al., 2025a). It consists of English texts annotated using Amazon Mechanical Turk, with binary labels (0 for absence, 1 for presence) assigned to five core emotions—anger, fear, joy, sadness, and surprise—alongside neutral instances. Given that multiple emotions can co-occur within a single text, this introduces an additional layer of complexity to the classification task.

The dataset is split into 2,768 training examples, 116 development examples, and 2,767 test examples. The shortest text in the training set contains 4 tokens, while the longest extends to 121 tokens, demonstrating a diverse range of text lengths.

Figure 3 provides a pie chart illustrating the distribution of single-label examples across the five emotions and neutral category. To further explore multi-label patterns, Figure 4 presents a bar plot depicting the distribution of instances with varying numbers of assigned emotion labels. This visualization helps highlight the frequency and complexity of multi-label cases within the dataset.

4.2 Results

We evaluate the performance of RoBERTa-large, EmoT5, and our ensemble model, RoEmo, on both the development and test datasets. The results on the development set are presented in Table 1 and Table 2. As shown, RoEmo achieves the highest Macro-F1 score among all models, demonstrating its effectiveness in multi-label emotion classification.

Emotion	RoBERTa	EmoT5	RoEmo
Anger	0.7333	0.7879	0.7879
Fear	0.8201	0.7947	0.7821
Joy	0.7636	0.8276	0.8276
Sadness	0.7324	0.7356	0.7416
Surprise	0.7077	0.6667	0.7714

Table 1: Evaluation scores on the Dev set by emotion

	Macro F1	Micro F1
RoBERTa	0.7514	0.7667
EmoT5	0.7625	0.7653
RoEmo	0.7821	0.7783

Table 2: Evaluation scores on the Dev dataset

Emotion	RoBERTa	EmoT5	RoEmo
Anger	0.6553	0.6493	0.6746
Fear	0.8449	0.8344	0.8561
Joy	0.7666	0.7704	0.7654
Sadness	0.7699	0.7559	0.7861
Surprise	0.7391	0.7249	0.7473

Table 3: Evaluation scores on the Test set by emotion

Similarly, we evaluate the models on the test dataset, as shown in Table 3 and Table 4. While RoBERTa outperforms EmoT5 on test data, RoEmo achieves the highest Macro-F1 score, further confirming the effectiveness of our ensemble approach.

Overall, our results demonstrate that RoEmo consistently outperforms the individual models in terms of Macro-F1 score, making it a more effec-

	Macro F1	Micro F1
RoBERTa	0.7552	0.7837
EmoT5	0.747	0.7741
RoEmo	0.7659	0.7913

Table 4: Evaluation scores on the Test dataset

tive approach for multi-label emotion classification.

5 Conclusion

In this work, we tackled multi-label emotion detection in short texts by leveraging a hybrid approach that integrates both generative and discriminative models. By combining the instruction fine-tuned generative model with the of a discriminative model, our method effectively captures diverse emotional expressions while maintaining computational efficiency. Through an ensemble strategy, we demonstrated that merging predictions from both models enhances classification performance. Our analysis of recent developments in the field, along with empirical results on SemEval-2025 Task 11, highlights the effectiveness of our approach. The findings suggest that this hybrid framework is a promising direction for improving emotion detection systems, and balancing generalization and efficiency.

Limitations

Despite the strong performance of our approach, RoEmo has some limitations. It relies on two models—RoBERTa-Large and EmoT5—without explicitly capturing relationships between emotions, which could enhance contextual understanding. The ensemble setup also increases computational cost and complexity, limiting its practicality in real-time or resource-constrained settings. Additionally, we did not compare against more recent LLMs such as Llama 3 (Grattafiori et al., 2024) or Qwen 2.5 (Qwen et al., 2025). Future work could explore other baselines, investigate alternative fusion methods, address data imbalance, and develop more efficient architectures that model inter-emotion dependencies while reducing computational overhead.

Acknowledgments

We sincerely appreciate Dr. Kristen Johnson and her group at Michigan State University for their invaluable support and insightful feedback.

References

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Firoj Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*.

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Venkatesh Duppada, Royal Jain, and Sushant Hiray. 2018. SeerNet at SemEval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 18–23, New Orleans, Louisiana. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar Zaïane. 2021. Seq2Emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4717–4724, Online. Association for Computational Linguistics.

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*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1–10.

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.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

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.

Jianhui Pang, Yanghui Rao, Haoran Xie, Xizhao Wang, Fu Lee Wang, Tak-Lam Wong, and Qing Li. 2021. Fast supervised topic models for short text emotion detection. *IEEE Transactions on Cybernetics*, 51(2):815–828.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85:333–359.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Armin Seyeditabari, Narges Tabari, and Wlodek Zadrozny. 2018. Emotion detection in text: a review. *Preprint*, arXiv:1806.00674.

Peng Xu, Zihan Liu, Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2020. Emograph: Capturing emotion correlations using graph networks. *CoRR*, abs/2008.09378.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium. Association for Computational Linguistics.