RSSN at SemEval-2025 Task 11: Optimizing Multi-Label Emotion Detection with Transformer-Based Models and Threshold Tuning

Ravindran V Rajalakshmi S Angel Deborah S
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai - 603110, Tamil Nadu, India
{ravindran2213003, rajalakshmis, angeldeborahS}@ssn.edu.in

Abstract

Multi-label emotion classification in NLP requires models to capture complex emotional nuances in text. This study explores transformer-based models, primarily fine-tuning *BERT-base-uncased*, for classifying five perceived emotions: *anger, fear, joy, sadness*, and *surprise*. As part of SemEval 2025 Task 11 (Track A) in English, we preprocess text using tokenization, stopword removal, and lemmatization. Baseline models employing logistic regression with TF-IDF establish performance benchmarks. To address class imbalance, we fine-tune BERT using weighted binary cross-entropy loss, further improving classification with threshold optimization. Experimental results demonstrate that fine-tuned BERT significantly outperforms traditional approaches, achieving a macro F1-score of 0.6675, which rises to 0.7062 after threshold optimization. Comparative analysis against RoBERTa fine-tuning, CNN-TF-IDF hybrids, and XGBoost classifiers highlights the superiority of contextual embeddings for multi-label classification. While threshold tuning enhances recall and precision, challenges like class imbalance and inter-class confusion persist, motivating future research into ensemble models and domain-adaptive training.

1 Introduction

Emotion classification in NLP is essential for identifying *perceived emotions*, which reflect how an audience interprets a speaker's sentiment. Unlike sentiment analysis, which categorizes text as positive, negative, or neutral, multi-label emotion detection captures multiple co-occurring emotions in a single instance.

This study focuses on SemEval 2025 (Track A) for multi-label emotion detection in English, classifying text snippets into six perceived emotions: *joy, sadness, anger, fear, surprise*, and *disgust*. Rather than identifying the speaker's true emotions or reader's reactions, the task centers on commonly inferred emotions, influenced by linguistic and cultural factors.

We fine-tune a BERT-based model to capture contextual dependencies while addressing class imbalance and label co-occurrence challenges. Our approach includes text preprocessing with SpaCy, dataset analysis, and hyperparameter tuning. TF-IDF-based models serve as baselines, and weighted binary cross-entropy loss is used for training. Performance is evaluated via F1-score, the official metric for the task.

Experimental results show that transformer-based models significantly outperform traditional methods, effectively detecting multiple emotions per instance. Despite improvements, challenges like inter-class confusion and label ambiguity remain. This study provides insights into optimizing multi-label emotion classification and outlines directions for future research.

2 Related Work

CM-MEC-21, introduced by Ameer et al. (Tang et al., 2020), serves as a benchmark dataset for multi-label emotion classification in code-mixed (English-Roman Urdu) SMS messages. The study evaluated traditional machine learning, deep learning, and transformer-based models (BERT, XLNet), revealing that n-gram-based features with OVR Naïve Bayes achieved the highest Micro-F1 score of 0.67. This finding underscores the limitations of deep learning in low-resource environments and the continued relevance of feature-driven approaches.

Exploring subjectivity and sentiment analysis, Wiebe et al. (Wiebe et al., 2005) introduced a detailed corpus annotation framework. Their methodology focused on phrase-level subjectivity rather than sentence-level labels, incorporating

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 773–779

nuanced elements such as private states, beliefs, and nested sources of emotion. The framework has since been widely adopted for opinion mining and sentiment classification tasks.

To enhance emotion classification from text, Abas et al. (Abas et al., 2022) proposed a hybrid model combining BERT embeddings with a CNN-based classifier. Their approach leveraged BERT's contextual word representations while utilizing CNNs for final classification. Evaluations on the SemEval-2019 Task 3 and ISEAR datasets demonstrated that the BERT-CNN model outperformed baseline methods, achieving an F1-score of 94% on SemEval and 76% on ISEAR.

Shifting towards non-transformer-based approaches, Liu et al. (Liu et al., 2023) refined the multi-label K-Nearest Neighbors (MLkNN) algorithm for short-text emotion classification. Their model incorporated both local sentence-level features and global contextual dependencies, improving classification accuracy through iterative refinement based on emotion transfer probabilities. Experiments on the Sentiment140 Twitter corpus demonstrated that the enhanced MLkNN model (with optimized $K = 8$ and $\alpha = 0.7$) outperformed traditional MLkNN approaches, achieving a recall rate of 0.8019. Their findings highlight the effectiveness of integrating local and global contextual information for multi-label emotion classification.

3 System Overview

3.1 Data Preprocessing and Exploration

The dataset (Muhammad et al., 2025a) used in this study consists of short text samples annotated with multiple emotion labels. The emotions considered are *anger, fear, joy, sadness, and surprise*. Each text instance may be associated with one or more emotion labels, making it a multi-label classification task. The dataset was loaded and analyzed to understand its structure and characteristics.

3.1.1 Preprocessing Steps

To ensure data quality and enhance feature extraction for the classification model, a series of preprocessing steps were applied:

- **Text Normalization:** All text was converted to lowercase, and punctuation and numerical characters were removed.
- **Stopword Removal:** Non-informative words were filtered using the built-in SpaCy stopword list.
- **Lemmatization:** Words were reduced to their base forms using SpaCy's lemmatizer to standardize textual representations.
- **Negation Handling:** Words following negation terms such as "not", "never", and "no" were concatenated with the negation marker (e.g., *not good* $\rightarrow$ *not_good*).
- **Rare and Frequent Word Removal:** Words appearing with extremely low frequency (less than 2 occurrences) or extremely high frequency (above 95% of total words) were removed to mitigate noise.

After preprocessing, the cleaned text was saved for further analysis and model training.

3.1.2 Dataset Exploration and Visualization

To understand the dataset distribution, various statistical and visual analyses were performed.

Text Length and Word Count Distribution To analyze textual characteristics, the distribution of text lengths (character count) and word counts was visualized. Figures 1 and 2 illustrate these distributions.

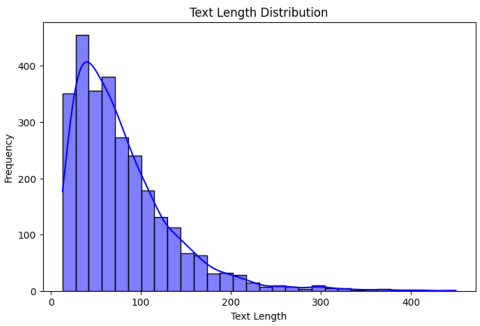

Figure 1: Text Length Distribution

Emotion Correlation Analysis To examine relationships between different emotions, a correlation matrix was computed using the label co-occurrence data. Figure 3 presents the heatmap of correlation values.

The dataset is imbalanced, with *fear* as the most frequent label. Many instances contain multiple emotions, requiring a robust multi-label approach. Most texts are under 100 characters, and

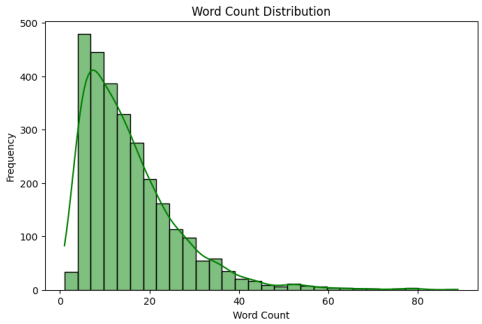

Figure 2: Word Count Distribution

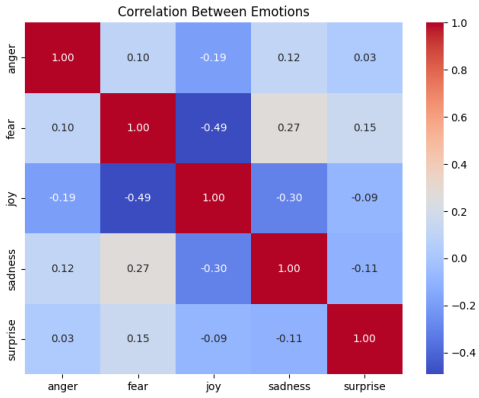

Figure 3: Correlation Between Emotions

co-occurrence analysis highlights strong associations between specific emotion pairs, aiding label dependency modeling.

3.2 Threshold Optimization Methodology

For multi-label classification, we optimize prediction thresholds for each emotion class to maximize F1-scores. The optimization process begins by computing the class probabilities from the model logits using a sigmoid activation function. We then evaluate precision, recall, and F1-scores across a range of thresholds, $\tau \in [0.1, 0.9]$, with a step size of 0.01. For each emotion class, we select the optimal threshold, τ_c^*, which maximizes the F1-score for that class. Specifically, the optimal threshold is determined by:

$$\tau_c^* = \tau \in [0.1, 0.9] \text{argmax} \frac{2 \cdot P_c(\tau) \cdot R_c(\tau)}{P_c(\tau) + R_c(\tau)},$$

where $P_c(\tau)$ and $R_c(\tau)$ represent the precision and recall for class c at threshold τ. Once the optimal thresholds are determined, they are applied during inference to maximize the performance of the classification model.

In particular, for *anger*, lowering the threshold from 0.5 to 0.21 resulted in a significant improvement in recall by 21.5% (from 0.354 to 0.569), at a modest cost to precision (from 0.697 to 0.617). This adjustment helped mitigate the issue of under-detection in *anger* cases. The threshold optimization also highlighted the different characteristics of each emotion class. For *fear*, the optimal threshold was higher, at 0.33, which reflected the confident predictions made for this emotion. In contrast, *sadness* benefited from a lower threshold of 0.21, better capturing its more subtle expressions.

While threshold optimization provided overall performance gains, it also introduced some trade-offs. One such trade-off was the impact on precision, especially for minority classes like *anger*, where the precision dropped by 8% in order to achieve a 21.5% increase in recall. Additionally, false positives increased for some of the minority classes, although this was mitigated by the significant reduction in false negatives, resulting in a 7% overall increase in recall.

The threshold optimization process led to several noteworthy improvements. Specifically, the macro F1-score increased from 0.6435 to 0.6814, reflecting a 5.88% improvement. The most dramatic gain was observed in *anger*, which saw an impressive 26.1% increase in its F1-score. Other emotion classes, such as *joy* and *surprise*, also saw balanced improvements, with their F1-scores increasing by 10.4% and 10.2%, respectively. These results are summarized in Table 1.

4 Baseline Approaches

To establish a foundational benchmark for multi-label emotion classification, we experimented with logistic regression models trained on TF-IDF representations of text. These models were selected due to their efficiency and interpretability in text classification tasks. Multiple variations were explored to examine the impact of feature engineering, class imbalance handling, and threshold optimization.

4.1 TF-IDF with Logistic Regression

The first model utilized a TF-IDF vectorizer with a vocabulary size of 5000, setting a document frequency range of 2% to 90%. A One-vs-Rest logistic regression classifier was trained on these features. The training and evaluation were conducted using an 80-20 train-test split.

4.1.1 Logistic Regression with Class Weights

To mitigate the issue of class imbalance, a weighted logistic regression model was trained using class

Table 1: Threshold Optimization Ablation Study

Emotion	Optimal τ	F1 (0.5)	F1 (Opt)	Imp.	Prec (0.5)	Prec (Opt)	Rec (0.5)	Rec (Opt)
Anger	0.21	0.469	0.592	+26.1%	0.697	0.617	0.354	0.569
Fear	0.33	0.828	0.853	+3.0%	0.848	0.823	0.809	0.885
Joy	0.28	0.633	0.699	+10.4%	0.728	0.696	0.560	0.701
Sadness	0.21	0.669	0.709	+6.1%	0.789	0.665	0.581	0.760
Surprise	0.24	0.625	0.689	+10.2%	0.741	0.681	0.541	0.698

weights computed dynamically based on label distributions. This approach aimed to improve recall for minority classes while maintaining precision for dominant classes.

4.1.2 Feature Engineering and Threshold Optimization

Refinements included expanding TF-IDF to 10,000 tokens with bigrams for better context, applying sublinear scaling to balance term frequencies, and optimizing thresholds to enhance classification performance.

While these refinements contributed to an increase in overall performance, the results highlight the limitations of traditional models in capturing nuanced emotional expressions. These findings motivate the need for more sophisticated representations, such as deep contextual embeddings, to better model complex emotional variations.

5 Advanced Implementations and Experimental Analysis

To tackle multi-label emotion classification, we explored various deep learning approaches, leveraging pre-trained transformers and hybrid architectures integrating TF-IDF and CNNs.

5.1 BERT Fine-Tuning for Multi-Label Emotion Classification

We fine-tuned *BERT-base-uncased* for multi-label classification using tokenized input (max sequence length = 128). Training employed the AdamW optimizer ($2e^{-5}$ learning rate, 0.01 weight decay) with binary cross-entropy loss. While achieving strong performance, particularly in detecting *fear* (F1 = 0.79), the model struggled with *anger* (F1 = 0.46). Threshold optimization improved macro F1-score from 0.6491 to 0.6816.

5.2 RoBERTa Fine-Tuning for Multi-Label Emotion Classification

Using *RoBERTa-base*, we followed a similar fine-tuning approach but with a higher learning rate ($3e^{-5}$) and a 10% warm-up proportion. Initially, it failed to learn representations for most emotions, resulting in a low macro F1-score (0.1493). After threshold optimization, performance improved significantly (macro F1-score = 0.4547), though classification imbalance remained a challenge.

5.3 DistilBERT Embeddings with XGBoost Classifier

DistilBERT's `[CLS]` token embeddings were extracted and used as input for an *XGBoost* classifier (50 estimators, depth 4, learning rate 0.1). Although it leveraged contextual embeddings, it did not match fine-tuned transformers, achieving a macro F1-score of 0.4671 with poor recall for *anger* (F1 = 0.15).

5.4 RoBERTa Embeddings with XGBoost Classifier

RoBERTa embeddings were extracted similarly and trained with XGBoost (100 estimators, 0.05 learning rate). However, its macro F1-score (0.2472) indicated that static embeddings alone were insufficient for effective classification, leading to poor recall for most emotions except *fear*.

5.5 CNN with TF-IDF and DistilBERT Hybrid Model

We combined TF-IDF features (10,000 n-grams) with DistilBERT embeddings in a CNN architecture using convolutional layers, max-pooling, and dropout. The hybrid approach improved representation learning but remained limited by dataset constraints, with a macro F1-score of 0.5125 and relatively weak performance for *joy* and *sadness*.

Fine-tuned BERT and RoBERTa models outperformed other approaches, demonstrating the effectiveness of contextual embeddings. Threshold optimization improved recall, while XGBoost with transformer embeddings failed to capture deep dependencies. CNN-hybrid models leveraged multi-feature representation but remained limited by data constraints.

5.6 Experimental Evaluation and Test Results

We validated our approaches on an external test set using two fine-tuned *BERT-base-uncased* models with different learning rates to assess generalization, optimize classification thresholds, and refine fine-tuning strategies. For all models evaluated in this study, including logistic regression, transformer-based models, and hybrid approaches, we maintained consistency by using the same validation set for threshold optimization and the same held-out test set for final performance evaluation, ensuring fair comparison.

5.6.1 Fine-Tuned BERT Models

The first model was trained with a learning rate of $3e^{-5}$, batch size 16, and a warm-up proportion of 10%, applying early stopping over 10 epochs. The second model used a lower learning rate of $2e^{-5}$ for improved stability. Both models employed AdamW optimization and binary cross-entropy loss for multi-label classification.

Results and Threshold Optimization Before threshold optimization, both models achieved a macro F1-score of approximately 0.667, with *fear* consistently scoring highest and *anger* the lowest. Applying optimized classification thresholds improved the macro F1-score to 0.7047 for Model 1 and 0.7062 for Model 2, enhancing recall, particularly for underrepresented emotions. The results indicate that threshold tuning significantly refines decision boundaries, and variations in learning rates have minimal impact when proper threshold selection is applied.

5.6.2 Final Predictions and Observations

Using the best-performing model (learning rate $2e^{-5}$ with optimized thresholds), predictions on the external test set showed stable F1-scores across all emotions. The model effectively generalized, with threshold tuning improving recall and classification accuracy. However, class imbalance persisted, suggesting potential enhancements through ensemble learning and data augmentation.

Fine-tuned BERT models, combined with threshold optimization, demonstrated superior multi-label classification performance. Learning rate variations had little impact on final results when proper threshold tuning was applied. The study underscores the necessity of threshold optimization for imbalanced datasets, proving the effectiveness of transformer-based fine-tuning for emotion classification in real-world applications.

6 Results and Performance Evaluation

This section presents a comprehensive analysis of the results obtained from our experiments on multi-label emotion classification. The models were fine-tuned on the preprocessed dataset and evaluated based on key performance metrics, including macro F1-score, precision, recall, and accuracy. Additionally, we conducted threshold optimization to enhance the classification performance. The results are consolidated in the following subsections.

6.1 Initial Performance Evaluation

The logistic regression model trained on TF-IDF features exhibited strong performance for frequent emotions (*fear*, *surprise*) but struggled with minority classes (*anger*, *joy*). Introducing class weights improved recall, and further threshold tuning enhanced classification, achieving a macro F1-score of 0.54, as shown in Table 2.

Table 2: Performance of Logistic Regression Models with TF-IDF

Emotion Class	Initial Model	With Class Weights	Optimized Model
Anger	0.03	0.33	0.35
Fear	0.73	0.67	0.76
Joy	0.17	0.42	0.44
Sadness	0.38	0.54	0.58
Surprise	0.43	0.61	0.65
Macro F1-score	0.32	0.50	0.54

The initial evaluation of the fine-tuned *BERT-base-uncased* models, before applying threshold optimization, revealed significant variations in classification performance across different emotion classes. Table 3 summarizes the macro F1-score and per-class F1-scores before threshold optimization.

Table 3: Performance of Fine-Tuned BERT Models Before Threshold Optimization

Emotion Class	Model 1 (LR = $3e^{-5}$)	Model 2 (LR = $2e^{-5}$)
Anger	0.53	0.54
Fear	0.81	0.80
Joy	0.65	0.64
Sadness	0.65	0.67
Surprise	0.71	0.68
Macro F1-score	0.6678	0.6675

The results indicate that *fear* was consistently the best-classified emotion, achieving an F1-score above 0.80 in both models, suggesting that the model effectively captured its contextual cues. In contrast, *anger* had the lowest performance, highlighting the difficulty in distinguishing it from other

emotions in multi-label classification. Additionally, the macro F1-scores of both models remained nearly identical before threshold optimization, indicating that variations in learning rates had minimal impact on classification performance.

6.2 Impact of Threshold Optimization

To refine the predictions and improve model performance, we optimized classification thresholds for each emotion class. Instead of using the default threshold of 0.5, an optimal probability threshold was determined by maximizing the per-class F1-score. The performance gains achieved through this optimization are presented in Table 4.

Table 4: Performance of Fine-Tuned BERT Models After Threshold Optimization

Emotion Class	**Model 1 (LR = $3e^{-5}$)**	**Model 2 (LR = $2e^{-5}$)**
Anger	0.60	0.61
Fear	0.82	0.82
Joy	0.68	0.65
Sadness	0.69	0.69
Surprise	0.74	0.76
Optimized Macro F1-score	0.7047	0.7062

Threshold optimization proved highly effective, increasing the macro F1-score by approximately 3–4%. The most significant improvement was observed in *anger*, where the F1-score rose from 0.53 to 0.61, addressing its previously weak performance. *Surprise* benefited the most, with an F1-score increase of 4–8%, enhancing recall while maintaining precision. Notably, the impact of learning rate variations remained minimal after optimization, with Model 2 showing a slight edge in macro F1-score.

6.3 Performance on Test Set

The final evaluation was conducted on an unseen test set using the best-performing model (Model 2, learning rate = $2e^{-5}$, optimized thresholds). Table 5 presents the final performance metrics.

Table 5: Final Performance on External Test Set

Metric	**Before Optimization**	**After Optimization**
Macro F1-score	0.6675	0.7062
Micro F1-score	0.72	0.75
Weighted F1-score	0.71	0.74
Precision	0.70	0.73
Recall	0.69	0.76

The external test set evaluation confirmed the model's strong generalization, as the macro F1-score remained consistent. Notable improvements were observed in micro and weighted F1-scores, indicating enhanced prediction stability across all emotion classes. Additionally, recall increased by 7% post-optimization, demonstrating that threshold tuning effectively reduced false negatives and improved overall classification performance.

6.4 Conclusion

Fine-tuned BERT models proved highly effective for multi-label emotion classification, outperforming traditional methods like logistic regression with TF-IDF. Threshold optimization significantly improved recall and decision boundaries, especially for underrepresented emotions like anger and joy.

While fear and surprise were well-classified, anger remained the most challenging, highlighting difficulties in distinguishing subtle emotional cues. Alternative models including XGBoost with transformer embeddings fell short, emphasizing the importance of contextualized embeddings. Challenges like class imbalance and inter-class confusion persist, suggesting future work on ensemble learning, contrastive pretraining, and domain-adaptive fine-tuning. Integrating textual, visual, and audio cues through multi-modal approaches could further improve real-world emotion detection.

References

Ahmed R Abas, Ibrahim Elhenawy, Mahinda Zidan, and Mahmoud Othman. 2022. Bert-cnn: A deep learning model for detecting emotions from text. *Computers, Materials & Continua*, 71(2).

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)*, pages 117–121. IEEE.

S. Angel Deborah, S. Rajalakshmi, S. Milton Rajendram, and T. T. Mirnalinee. 2020. Contextual emotion detection in text using ensemble learning. In *Emerging Trends in Computing and Expert Technology*, pages 1179–1186, Cham. Springer International Publishing.

Diogo Cortiz. 2021. Exploring transformers in emotion recognition: a comparison of bert, distillbert, roberta, xlnet and electra. *arXiv preprint arXiv:2104.02041*.

Xuan Liu, Tianyi Shi, Guohui Zhou, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Emotion classification for short texts: an improved multi-label method. *Humanities and Social Sciences Communications*, 10(1):1–9.

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.

Angel Deborah S, Rajalakshmi S, S Milton Rajendram, and Mirnalinee T T. 2018. SSN MLRG1 at SemEval-2018 task 1: Emotion and sentiment intensity detection using rule based feature selection. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 324–328, New Orleans, Louisiana. Association for Computational Linguistics.

Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. 2020. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256.

Ravindran V, Shreejith Babu G, Aashika Jetti, Rajalakshmi Sivanaiah, Angel Deborah, Mirnalinee Thankanadar, and Milton R S. 2024. TECHSSN at SemEval-2024 task 10: LSTM-based approach for emotion detection in multilingual code-mixed conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 763–769, Mexico City, Mexico. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.