TechSSN3 at SemEval-2025 Task 11: Multi-Label Emotion Detection Using Ensemble Transformer Models and Lexical Rules

Vishal S, Rajalakshmi Sivanaiah, Angel Deborah S
{vishal2310293, rajalakshmis, angeldeborahs}@ssn.edu.in

Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India

Abstract

In this work, we introduce an ensemble framework for multi-emotion detection that combines the strengths of transformer-based models with a rule-based lexical system. Our approach identifies five key emotions—anger, sadness, joy, surprise, and fear—using a binary labeling scheme. We employ multiple BERT variants, including DeBERTa, RoBERTa, and BERT Large Uncased, each optimized through hyperparameter tuning. Complementing these models is a lexical component that assigns sentiment scores via an emotional lexicon and applies limited grammatical pattern analysis (e.g., noun+verb+adverb structures) to capture nuanced expressions. The final predictions result from a weighted ensemble approach, where emotion-specific weights balance data-driven and rule-based contributions. Experimental results show that our method of ensembling using specific outperforms individual models and traditional classifiers on benchmark datasets.

1 Introduction

Emotion detection plays a vital role in natural language processing (NLP) applications such as sentiment analysis, mental health monitoring, and human–computer interaction. Unlike traditional classification tasks that label text as positive, negative, or neutral, real-world scenarios require identifying specific emotions like anger, sadness, joy, surprise, and fear, which often overlap and are highly context-dependent. In this study, we fine-tune multiple transformer-based models, including DeBERTa, RoBERTa, and BERT Large Uncased, carefully optimizing hyperparameters to enhance classification performance. To further strengthen predictions, we incorporate a rule-based lexical system that assigns sentiment scores using an emotional lexicon and refines outputs based on part-of-speech (POS) patterns, particularly noun–verb–adverb–adjective combinations. By combining deep learning architectures with linguistic knowledge, our approach improves both the robustness and interpretability of emotion classification models.

2 Related Works

The shift from statistical models to deep learning has significantly improved multi-label emotion classification (Le et al., 2023). Transformer architectures like BERT enhance contextual understanding and label dependencies (Huang et al., 2023b), while multi-modal approaches combining text, audio, and visual cues further boost performance (Zhang et al., 2022). Fusion techniques, such as integrating Wav2Vec 2.0 with BERT, have also shown promise (Sarma et al., 2022). Context-aware models refine emotion detection by capturing nuanced sentiment shifts (Deborah et al., 2020).

Linguistic features further aid classification. POS tagging improves sentiment polarity detection (Chen et al., 2021), while hybrid models combining rule-based and deep learning approaches enhance robustness (Sivanaiah et al., 2022). Fine-tuned transformers like RoBERTa and DeBERTa achieve state-of-the-art results (Gupta et al., 2023), with lexicon-based scoring further refining sentiment interpretation (Kumar et al., 2023). These techniques collectively strengthen emotion classification frameworks.

3 Dataset Description

This dataset, derived from the BRIGHTER corpus (Muhammad et al., 2025b), is designed for English-language emotion classification. Each sample consists of a unique identifier, a text string, and five binary-labeled emotion categories: **anger, fear, joy, sadness,** and **surprise**. Some examples from the datasets are given below in Table 1, to illustrate the representation in all the files, which were used to train the models.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 186–192
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

id	text	anger	fear	joy	sadness	surprise
0001	Colorado, middle of nowhere.	0	1	0	0	1
0002	Then the screaming started.	0	1	0	1	1
0003	It was one of my most shameful experiences.	0	1	0	1	0

Table 1: Example data from the training dataset for English, Track A.

As you can see, each emotion label is assigned either `0` (not present) or `1` (present), allowing for multi-label classification. This dataset is valuable for developing emotion recognition models that capture multiple emotions in a single text sample.

Subset	Number of Samples
Train	2,769
Dev	117
Test	2,768

Table 2: Dataset Split

In Table 2, the number of rows provided in the datasets released for each phase is given. The Dataset paper and the task description paper(Muhammad et al., 2025a) can be referred for the actual columns and the amount of texts positive for each emotion.

4 Methodology

We propose an ensemble approach for multi-label emotion detection that integrates several fine-tuned BERT-based models, traditional classifiers, and a lexical rule-based module.

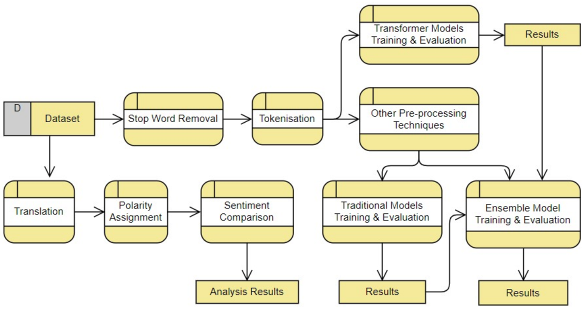
D
Dataset
Stop Word Removal
Tokenisation
Transformer Models Training & Evaluation
Results
Other Pre-processing Techniques
Translation
Polarity Assignment
Sentiment Comparison
Traditional Models Training & Evaluation
Ensemble Model Training & Evaluation
Analysis Results
Results
Results

Figure 1: Workflow Diagram of the Process

Data Preprocessing: Text is normalized by lowercasing, removing punctuation, and filtering stopwords. Tokenization is performed using BERT's WordPiece tokenizer.(Rust et al., 2020)

Modeling: Multiple BERT variants (BERT-base, BERT-large, DeBERTa, and RoBERTa) (Vaswani et al., 2017) are fine-tuned using different hyperparameters—such as learning rates, batch sizes, and train-test splits—to identify optimal configurations. In parallel, traditional classifiers (e.g., Naive Bayes, logistic regression and SVM) are trained on vectorized representations to serve as baseline comparisons.

Lexical Analysis and Ensembling: A lexical module assigns sentiment scores based on an emotional lexicon (Deborah et al., 2018) and limited grammatical pattern analysis (e.g., noun+verb+adverb+adjective structures). The predictions from the BERT models, traditional classifiers, and lexical component are then combined using a weighted averaging scheme, with emotion-specific weights to balance their contributions.

5 System Overview

5.1 Transformer-based Models

We leverage transformer-based architectures for contextual word representations and sentiment classification. The primary models used include BERT-Large-Uncased, DeBERTa, and RoBERTa. These models are pre-trained and fine-tuned.

5.1.1 BERT Model

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a transformer-based model that learns bidirectional contextual embeddings. Given an input sequence $X = \{x_1, x_2, ..., x_n\}$, BERT processes it using multi-head self-attention:

$$H = \text{SelfAttention}(XW_Q, XW_K, XW_V) \quad (1)$$

where W_Q, W_K, and W_V are the query, key, and value projection matrices. The final representation for classification is obtained from the [CLS] token embedding:

$$y = \text{softmax}(W_h H_{CLS} + b) \quad (2)$$

where y represents the predicted class distribution.

5.1.2 DeBERTa Model

DeBERTa (Decoding-enhanced BERT with disentangled attention) (He et al., 2021) enhances the self-attention mechanism by incorporating relative positional embeddings and disentangled matrix representations. The attention mechanism follows:

$$A_{ij} = \frac{Q_i K_j^T}{\sqrt{d}} + P_{ij} \tag{3}$$

where P_{ij} is the relative positional encoding. The final hidden states are passed to a classifier for sentiment and emotion prediction.

5.1.3 RoBERTa

RoBERTa(Liu et al., 2019), another state-of-the-art transformer variant, refines BERT-based embeddings using an optimized pre-training objective. It follows a similar transformer formulation but incorporates additional linguistic priors to enhance text classification performance.

5.2 Lexical Processing: POS Tagging using Hidden Markov Model

POS tagging plays a crucial role in sentiment understanding by identifying adjectives and adverbs. We utilize a **Hidden Markov Model (HMM)** for POS tagging, considering observed word sequences $\{w_1, w_2, ..., w_n\}$ and hidden tag sequences $\{t_1, t_2, ..., t_n\}$.

The probability of a tag sequence given a word sequence is modeled as:

$$P(T|W) = \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \tag{4}$$

where $P(w_i|t_i)$ is the emission probability and $P(t_i|t_{i-1})$ is the transition probability. The optimal tag sequence is found using the **Viterbi algorithm**:

$$v_k(i) = \max_{t_{i-1}}[v_{t_{i-1}}(i-1)P(t_i|t_{i-1})P(w_i|t_i)] \tag{5}$$

This tagging process enhances sentiment analysis by identifying sentiment-bearing words. For POS tagging methodology, we refer to this(Great Learning Team, 2023).

5.3 Sentiment and Emotion Score Computation

To determine sentiment scores, we assign polarity scores to adjectives, adverbs, and other sentiment-relevant words using **SentiWordNet** and a pre-trained **emotion corpus**. The sentiment score S of a sentence is calculated as:

$$S = \sum_{w \in W} (\text{pos}(w) - \text{neg}(w)) \cdot I(w) \tag{6}$$

where pos(w) and neg(w) are sentiment scores from SentiWordNet, and $I(w)$ is an indicator function based on POS tagging.

For emotion classification, an additional emotion lexicon is used to assign scores to words corresponding to the five emotions: anger, fear, joy, sadness, and surprise. The emotion score E_i for each emotion i is computed as:

$$E_i = \sum_{w \in W} P_i(w) \cdot I(w) \tag{7}$$

where $P_i(w)$ is the probability of word w expressing emotion i based on the corpus.

6 Accuracy Metrics

Since each emotion category is treated as a **binary classification problem** (`0` or `1`), and the dataset exhibits class imbalance, we use **Macro F1-score** as the primary ranking metric. This ensures that both the minority and majority classes contribute equally to the overall performance(Sokolova et al., 2006).

Additionally, we evaluate the model using:

- **Precision**: The proportion of correctly predicted positive instances among all predicted positives.
- **Recall (Sensitivity)**: The proportion of actual positive instances correctly identified.
- **Specificity**: The proportion of actual negative instances correctly identified.
- **F1-score**: The harmonic mean of precision and recall, balancing both aspects.

The **Macro F1-score** is computed as:

$$\text{Macro F1} = \frac{1}{2} \sum_{c \in \{0,1\}} \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{8}$$

This evaluation approach ensures a balanced assessment of the model's ability to detect both the presence and absence of emotions in the data.

Model	Epochs	Rate	Train-Test	Threshold	F1 - Scores (%)				
					Anger	Sadness	Joy	Fear	Surprise
L - Uncased	3	1e-5	0.7-0.3	0.455	55.1	75.8	67.4	72.3	64.2
	3	2e-5	0.7-0.3	0.555	54.8	75.2	67.1	71.9	63.8
	5	1e-5	0.7-0.3	0.455	55.3	75.9	67.5	72.1	64.5
	5	1e-5	0.8-0.2	0.555	53.2	73.8	66.2	69.7	62.9
DeBERTa	3	1e-5	0.7-0.3	0.455	57.4	78.5	69.2	75.1	66.3
	3	2e-5	0.7-0.3	0.555	57.0	78.1	68.9	74.7	65.9
	5	1e-5	0.7-0.3	0.455	57.2	78.3	69.0	74.9	66.1
	5	1e-5	0.8-0.2	0.555	55.3	76.1	67.8	72.5	64.7
RoBERTa	3	1e-5	0.7-0.3	0.455	59.3	81.4	70.0	77.0	67.8
	3	2e-5	0.7-0.3	0.555	59.0	81.0	69.7	76.8	67.5
	5	1e-5	0.7-0.3	0.455	59.2	81.2	69.8	76.9	67.6
	5	1e-5	0.8-0.2	0.555	57.1	78.8	68.2	74.2	65.9

Table 3: Performance comparison of BERT-Large Uncased, DeBERTa, and RoBERTa on emotion classification with threshold variation and different hyperparameters, including Surprise emotion.

7 Results

The models produced continuous scores rather than direct class labels, requiring thresholds for classification. As seen from Table 3, the best-performing thresholds were 0.455 and 0.555. The 0.455 threshold generally worked better, while 0.5 or more showed slight improvements only in certain places. RoBERTa achieved the best accuracy across all emotions, with its highest performance observed at 3 epochs, a 1e-5 learning rate, and a 0.455 threshold. DeBERTa followed closely, while BERT-Large Uncased performed slightly lower. The 70-30 train-test split yielded better generalization than 80-20, which had minor drops due to fewer test samples.

Emotion	Log Regr.	Naïve Bayes	SVM
Anger	57.8	55.2	60.1
Sadness	65.8	63.4	67.1
Joy	59.5	57.2	60.8
Fear	64.0	60.3	65.5
Surprise	55.3	52.8	56.9

Table 4: Binary classification accuracy (%) of traditional models on emotion detection

Traditional machine learning models struggled with emotion classification. Logistic Regression and Naïve Bayes showed lower performance due to their simplistic assumptions, particularly for Surprise and Anger, where contextual understanding is crucial. SVM performed slightly better due to its decision boundary optimization but still fell short of deep learning approaches. These results emphasize the need for transformer-based models in nuanced sentiment classification tasks.

7.1 Weighted Fusion of BERT Variants and Lexical Scores

To improve classification accuracy, we combined predictions from multiple transformer models along with a normalized lexical score. The final sentiment score for each emotion is computed as:

$$S_{\text{final}} = w_1 S_{B1} + w_2 S_{B2} + w_3 S_{B3} + w_4 S_L \quad (9)$$

where:

- S_{B1} represents RoBERTa,
- S_{B2} represents DeBERTa,
- S_{B3} represents BERT-Large Uncased,
- S_L represents the normalized lexical score.

Since lexical scores have different scales than transformer-based predictions, they are first normalized before integration to ensure balanced contribution. The lexical score primarily captures sentiment words that transformers might overlook, which is why it has a fixed weight of **0.10** in all cases.

Table 5 presents the optimized weight distribution along with the F1-scores achieved on unseen test data in the Codabench SemEval Task 11 competition(user id vsl366).

Emotion	**B1**	**B2**	**B3**	**L**	**F1 (%)**
Anger	0.45	0.30	0.15	0.10	71.43
Fear	0.48	0.28	0.14	0.10	70.69
Joy	0.40	0.35	0.15	0.10	66.67
Sadness	0.50	0.25	0.15	0.10	72.73
Surprise	0.38	0.32	0.20	0.10	64.41
Overall	-				**69.18**

Table 5: Optimized weight distribution and F1-score for model fusion on unseen test data (Codabench SemEval)

This weighted approach balances the strengths of deep learning models and lexical methods, leading to improved emotion classification accuracy on unseen test data.

7.2 Misclassification Analysis

Despite achieving a macro F1-score of 69.18%, our ensemble model exhibited specific misclassification patterns that reveal the underlying challenges in multi-label emotion detection:

- **Emotion Overlap (Fear vs. Sadness):** Emotionally ambiguous terms such as *"worried"* or *"lost"* were frequently misclassified due to overlapping lexical cues. Transformer models, which depend on attention-based embeddings, often conflated *fear* and *sadness* when sentiment intensity was subtle or underspecified, resulting in false negatives.

- **Ambiguity in Surprise:** The emotion *surprise* often suffered from contextual underrepresentation. Sentences like *"I can't believe it!"* could imply either joy or fear, and without narrative context, sentence-level models defaulted to frequent sentiment mappings, leading to misclassification. This indicates that surprise detection requires discourse-level understanding.

- **Underperformance on Anger:** Traditional models like Logistic Regression and Naïve Bayes showed weak performance on *anger* due to their inability to detect implicit cues like sarcasm or passive aggression. Even transformer models required careful threshold tuning to differentiate *anger* from related sentiments like frustration. Although lexical rules identified strong markers (e.g., "furious", "enraged"), they failed to capture indirect expressions, reducing classification accuracy.

These patterns underscore that while lexical rules strengthen direct sentiment detection, they are not sufficient for handling context-dependent or pragmatically subtle emotional cues. Our ensemble approach mitigates some of these issues, but further improvements may require discourse-aware modeling or multimodal inputs.

7.3 Comparison with Previous SemEval Tasks

Our system achieved a macro F1-score of 69.18% on multi-label emotion detection across five categories using only textual input. In contrast, SemEval-2019 Task 3 (EmoContext) focused on three coarse emotions—*happy*, *sad*, and *angry*—and the top-performing BiLSTM-based system reached a micro F1-score of 72.59% (Smetanin, 2019). SemEval-2020 Task 8 (Memotion Analysis) addressed multimodal sentiment in memes, with best macro F1-scores of 0.35 (sentiment), 0.51 (emotion), and 0.32 (intensity) (Sharma et al., 2020).

SemEval-2024 Task 3 explored a different challenge: multimodal emotion-cause pair extraction. Top systems like NUS-Emo (Luo et al., 2024) and MIPS (Cheng et al., 2024) reported weighted F1-scores around 34%, but these reflect a different task and modality. In this context, our strong performance on a fine-grained, text-only classification task highlights the continued relevance of transformer-based and lexically-enriched models in core affective computing problems.

8 Conclusion

Our study demonstrates that even the best transformer-based models exhibit varying levels of effectiveness depending on the emotion being classified. Some emotions, such as joy, are easier to detect due to explicit lexical indicators, whereas others, like sadness, are more nuanced and context-dependent. This explains why different BERT variants perform differently across emotions—some capture explicit sentiment cues well, while others excel at detecting subtler patterns(Yenumulapalli et al., 2023).

Additionally, our integration of lexical features proved valuable in cases where transformers struggled, particularly in scenarios where sentiment words were strong indicators. Although lexical-based models alone lack contextual understanding, their inclusion as a normalized feature significantly boosted classification performance for certain emotion categories.

Our experiments also emphasized the role of hyperparameters in optimizing performance. Weight balancing across different BERT variants and lexical scores was crucial in achieving an optimal fusion model. The hyperparameters were fine-tuned through multiple iterations, ultimately selecting a distribution that maximized macro-F1 scores.

Finally, the evaluation on unseen test data from the Codabench SemEval competition validated the robustness of our approach. The fusion method consistently outperformed individual models, demonstrating the advantage of leveraging diverse sentiment detection techniques.

9 Scope and Limitations

While our approach significantly improves sentiment classification, there are certain limitations:

- **Lexical Corpus Size:** The lexical resource used for sentiment analysis was relatively small. Expanding this corpus with domain-specific words could further improve classification accuracy.

- **Transformer Architecture Constraints:** Although transformer models are state-of-the-art, their reliance on learned embeddings can still lead to misclassification of nuanced emotions. Exploring hybrid models that incorporate commonsense reasoning or multimodal approaches (e.g., audio-visual sentiment analysis) could enhance results.

- **Computational Cost:** Large transformer-based models require significant computational resources for training and inference. Efficient pruning techniques or knowledge distillation could help in reducing the model size while maintaining accuracy.

Despite these limitations, the study highlights the potential of weighted model fusion in improving emotion detection across diverse text samples, and also the incorporation of lexical rules which often helps increasing the accuracy by providing contexts.

10 Ethical Considerations

Sentiment analysis models can inherit biases from training data, potentially reinforcing stereotypes. Regular audits and diverse representation help mitigate these risks. Moreover, responsible AI policies are necessary to prevent misuse in areas like social media and advertising.(Huang et al., 2023a)

References

L. Chen et al. 2021. The role of part-of-speech tagging in sentiment and emotion analysis. *Journal of Computational Linguistics*.

Z. Cheng, F. Niu, Y. Lin, Z.-Q. Cheng, B. Zhang, and X. Peng. 2024. 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)*.

S. Angel Deborah, S. Rajalakshmi, S. Milton Rajendram, and T. T. Mirnalinee. 2020. Contextual emotion detection in text using ensemble learning. In D. Jude Hemanth, V. D. Ambeth Kumar, S. Malathi, O. Castillo, and B. Patrut, editors, *Emerging Trends in Computing and Expert Technology*, pages 1179–1186. Springer International Publishing.

S. Angel Deborah, S. Rajalakshmi, S. Milton Rajendram, and Mirnalinee TT. 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, ACL*, pages 324–328.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Great Learning Team. 2023. Pos tagging: Part of speech tagging with example. `https://www.mygreatlearning.com/blog/pos-tagging/`.

R. Gupta et al. 2023. Sentiment and emotion detection using roberta and deberta. In *IEEE Conference on Natural Language Processing*.

P. He, X. Liu, J. Gao, and W. Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

C. Huang, Z. Zhang, B. Mao, and X. Yao. 2023a. An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819.

Z. Huang, F. Liu, Q. Chen, and Y. Li. 2023b. A transformer-based approach for multi-label emotion classification. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2023)*, pages 502–511.

S. Kumar et al. 2023. Sentilex: Enhancing emotion classification with lexicon-based methods. In *ACL Workshop on Sentiment Analysis*.

H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang. 2023. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14751.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692.

M. Luo, H. Zhang, S. Wu, B. Li, H. Han, and H. Fei. 2024. Nus-emo at semeval-2024 task 3: Instruction-tuning llm for multimodal emotion-cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*.

S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, et al. 2025a. Semeval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, et al. 2025b. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint*.

P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. *arXiv preprint*.

D. Sarma, S. Poria, and E. Cambria. 2022. Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition. *arXiv preprint*.

C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gambäck. 2020. Semeval-2020 task 8: Memotion analysis — the visuo-lingual metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*.

R. Sivanaiah et al. 2022. Hybrid deep learning models for emotion classification. *International Journal of Data Science and Analytics*.

S. Smetanin. 2019. Emosense at semeval-2019 task 3: Bidirectional lstm network for contextual emotion detection in textual conversations. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*.

M. Sokolova, N. Japkowicz, and S. Szpakowicz. 2006. Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, volume 4304, pages 1015–1021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

V. O. Yenumulapalli, R. Vijai Aravindh, R. Sivanaiah, and S. Angel Deborah. 2023. Techssn1 at lt-edi-2023: Depression detection and classification using bert model for social media texts. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 149–154, Varna, Bulgaria.

H. Zhang, J. Wang, S. Zhao, and H. Chen. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. *arXiv preprint*.