

Improving Textual Ekman-Based Emotion Mapping on Social Media and RoBERTa Models

Khawla Hadi Nasserallah¹, Mohammed Hasan Abdulameer²

¹Department of Computer Science, Faculty of Education, University of Kufa, Iraq, khawlah.hassan@student.uokufa.edu.iq

²Department of Computer Science, Faculty of Education for women, University of Kufa, Iraq, : mohammed.almayali@uokufa.edu.iq

ARTICLE INFO

Article history:

Received: 15 /03/2026

Revised form: 02 /05/2026

Accepted : 03 /05/2026

Available online: 30 /09/2026

Keywords:: Emotion Recognition
Social Media Text, Deep Learning,
RoBERTa, GoEmotions Dataset

Abstract:

Recognizing emotions from social media short text continues to be a difficult problem due to the presence of serious class imbalance and semantic confusion in the fine-grained emotion corpus. Although the GoEmotions dataset is realistic, it involves 28 different emotion labels in an imbalanced manner, with significant class underrepresentation. Imbalance in class distribution negatively impacts the performance of multi-label classification models in terms of generalization capability and minority class detection.

In this paper, the effects of emotion label representations and adaptive decision criteria are investigated on multi-label emotion classification. More precisely, the number of emotion labels is reduced from 28 to seven basic emotions according to Ekman's theory to mitigate the problem of class imbalance and decrease the complexity of the classification problem. Moreover, a label-based threshold adjustment technique is used, where different optimal threshold values are determined.

In the experiments, we utilized a RoBERTa model that implements partial freezing of layers and make a comparison with a hybrid BERT-BiLSTM-Attention model within both fine-grained and Ekman-based configurations. The Binary Cross-Entropy loss function is used for training the model, while early stopping is implemented to avoid overfitting. In terms of results, our method provides an average Micro-F1 score of 0.6501 and Macro-F1 score of 0.5951 for the GoEmotions test dataset.

These results indicate that the reduction of the label set space positively affects the stability and accuracy of classification. Moreover, threshold tuning for individual labels allows balancing precision and recall and, consequently, achieves better results for minority classes. Therefore, label engineering and adaptive decisions prove to be more effective than complex models in improving emotion classification accuracy.

MSC..

<https://doi.org/10.29304/jqscsm.2026.18.32739>

1. Introduction

Emotion classification on social media texts has emerged as a new topic due to the increased amount of user-generated opinionated and emotional data in the online world. As opposed to sentiment analysis that considers polarity, emotion recognition helps in determining specific emotions like happiness, anger, and fear. This will give more insights about human activities, which will further enhance human-computer interaction [1].

Even with the advancement made in the field of emotion recognition, there still exists several challenges associated with short, informal social media text data. One of the major problems encountered in emotion classification is the class imbalance and semantic overlap that occur in fine-grained emotion databases, e.g., GoEmotions database, where 28 different emotions are found [2].

Approaches based on lexicons and classical machine learning techniques like n-gram and POS tagging feature, have proven to be ineffective for modeling complex emotions [3]. Models based on deep learning architectures, specifically CNNs, LSTMs, and BiLSTMs, performed better through learning of distributed representations of text [4]; however, they still struggle in terms of modeling dependencies at larger distances. Transformer models, including BERT and RoBERTa, have been

successful in emotion recognition tasks thanks to capturing context through attention mechanism [5], [6]

Nevertheless, model performance could not be attributed to architecture alone; rather, the design of the emotion label space might be equally important. The difficulty in fine-grained classification of emotions lies in their overlapping semantics and data sparsity issues.

To tackle these issues, theories of emotions based on psychological models such as Ekman's theory of basic emotions offer a rational approach to shrinking the number of labels used [7].

In this paper, we examine the influence of the model architecture and emotion label representations in multi-label emotion classification tasks. We compare two different approaches to building the model, namely the RoBERTa model and a hybrid model composed of BERT-BiLSTM-Attention, for the fine grained and Ekman-based settings. Moreover,

we employ label wise adaptive thresholding to mitigate the effect of imbalance in multi-label emotion classification problems.

The results show that decreasing the label space and using label-wise adaptive thresholds are crucial for achieving high performance in multi-label emotion classification problems.

Contributions of This Research:

This paper concentrates on multi-label emotion classification by proposing improvements in the areas of label design and decision-making that can effectively address class imbalance issues. The following contributions have been made in the process:

1. Emotion Label Engineering:

The fine-grained emotion labels are converted to fewer labels by leveraging Ekman's Theory on emotions, thereby solving the problem of class imbalance and overlapping semantics.

2. Model Architecture Comparisons:

A comparative evaluation of RoBERTa (transformer-based) and BERT-BiLSTM-Attention (hybrid) models has been done in the context of fine-grained and reduced emotion labels.

3. Label-wise Threshold Optimization:

Threshold optimization has been done label-wise to find the best decision boundary corresponding to each emotion label to ensure the right balance of precision and recall.

4. Comprehensive Evaluation:

The performance of the proposed approach has been analyzed from several perspectives, such as accuracy, F1 Score, precision, recall, and confusion matrix.

Unlike prior work focusing primarily on model architecture, this study emphasizes the role of label design and adaptive thresholding in improving classification performance.

2. Related Work

Text-based Emotion Recognition has been an area of research in natural language processing for over a decade and has evolved in terms of methods and methods in Emotion Representation. Traditionally, rule-based and lexicon-based methods, where a set of pre-defined dictionaries for emotion representation and linguistic features like n-gram, POS tagging, syntactic features, etc., were employed for emotion recognition in text data[8]. With advancements in machine learning, all the traditional machine learning classifiers, namely, SVM, Naive Bayes, and Logistic Regression, have become very sensitive to noise, domain variation, and informal language used in social media sites to such an extent that they have to be combined with traditional machine learning features, yet they still fail to comprehend the context and other forms of indirect expression of emotions, etc [9]. Deep learning was developed and it created a breakthrough in the study of emotion recognition. One of the first deep learning architectures, which was applied in sentence-level emotion classification, was the Convolutional Neural Networks (CNNs), which not only proved the effectiveness of distributed representations, but also confirmed the importance of automatic features, which were extracted during the process [10]. However, Recurrent Neural Networks (RNNs), particularly Long short-term memory (LSTM) and Bidirectional LSTM networks, gained much attention because of the ability to account for sequential dependencies and context in words, which was a major factor for many natural language processing tasks [11], [14]. In order to prevent this problem from occurring, the attention mechanism was included in the RNN-based architectures. The inclusion of this mechanism in RNN-based models will allow selective attention to words that are emotionally significant. This will also enhance the ability of the RNN-based models in the detection of implicit emotions in the text of social media postings [12]. Despite the enhancements that were included in the RNN-based models, these models faced limitations due to their sequence dependency.

New classes of transformer-based architectures were proposed, which revolutionized the field of text-based emotion recognition. Another class, BERT, presented the use of a self-attention model and applied contextual modeling in a bidirectional manner, registering state-of-the-art results in a large variety of NLP tasks, such as emotion classification [13]. Another variant, RoBERTa, which is based on BERT, applied optimization techniques to the pretraining step, resulting in the best generalization and robustness results [14]. A handful of recent works have also confirmed that RoBERTa outperforms other state-of-the-art transformers in emotion recognition, especially in large noisy social media datasets [15].

In addition, with the progress of development in the field of architecture, the problem of emotion representation scheme design is becoming more and more important. More detailed emotion datasets, like GoEmotions, provide detailed information related to data based on the 27 or 28 different categories of emotions, so the more precise emotional analysis is conducted [10]. However, many studies mentioned problems, such as the presence of a great

level of imbalance, sparsity, and semantic proximity between close emotions, which may lead to poor macro-level performance in generalization, as described in [16].

To address these issues, scholars have researched emotion space reduction methods grounded on psychological theories. In particular, the basic emotion model proposed by Ekman, which conceptualizes a limited number of universal emotions, is widely used since it is more interpretable and evidence-based [17], [18]. Some earlier works have suggested that down-sampling fine-grained labels of emotions to the categories used by Ekman can reduce noise when annotating, lower the class imbalance issue, and improve stability in classifications, especially for short-text cases [19], [20]. However, most such works are based on traditional classifiers or simple neural networks, and may not necessarily apply to state-of-the-art transformer-based models.

Yet, even after intensive research in the independent study of the transformer-based model and the emotion label, scant literature exists that can suffice to explore the interactivity between the two in the same framework of

evaluation. Further, it is not explored in depth whether the emotion-mapping technique as developed by Ekman can be used to boost the performance of even the state-of-the-art transformer-based models, such as the RoBERTa model. This is one of the motivating factors that encouraged the conducting of the research, which is a comparative analysis in depth in experimental scenarios [16].

Methodology

This study employs a controlled experimental design to investigate the impact of emotion classification representation and model structure on multi-category emotion classification, focusing on addressing category imbalances.

Figure 1 illustrates two paths: one for the hybrid model with 28 emotion types and the other for the model with 7 emotion types.

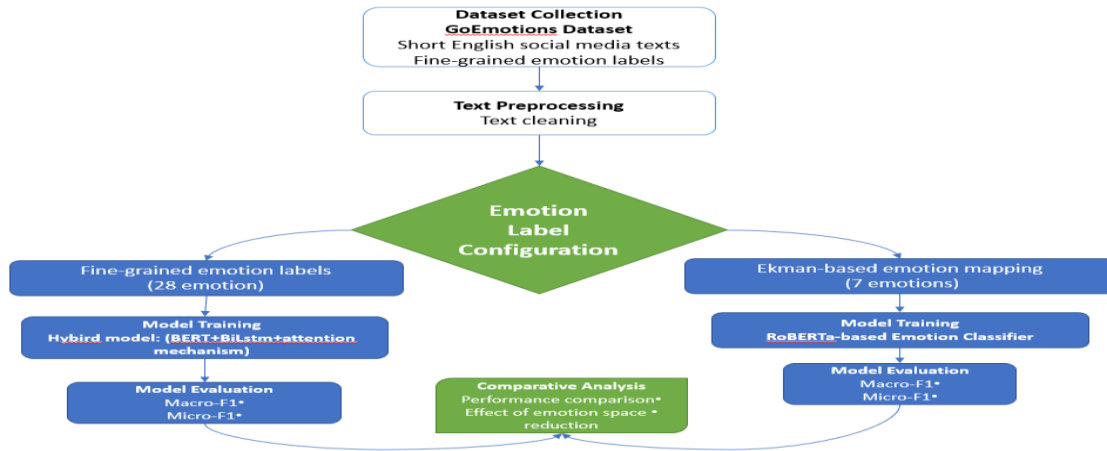


Figure 1: the proposed structure

3.1 Dataset

For the experiments, we use GoEmotions dataset [2] that consists of 54,260 social media comments written in English and labeled with multiple emotions. It is characterized by natural distribution of emotions and serious class imbalance when several emotions are much less frequent than others.

The pre-defined train, validation, and test sets for this research are shown in **Table 1** below.

Table 1. Dataset Splits

Split	Samples	Percentage
Training	43,408	80%
Validation	5,426	10%
Test	5,426	10%
Total	54,260	100%

The distribution of emotion tags in the corpus is highly skewed, as shown in Figure 2 below. Only a few types of emotions, including neutral, gratefulness, and happiness, are present in large numbers, while some others are extremely rare.

Figure 2: Frequency Distribution of Emotion Labels in GoEmotions Training Dataset

3.2 Data Preprocessing

A common set of pre-processing steps is used to transform the textual data into an appropriate form for model training. These steps include URL/user mentions removal, noisy text cleaning, lowercasing, and text tokenization with the RoBERTa tokenizer.

Punctuation and emotive symbols are retained as they might contain useful information on emotions expressed by users in social media posts.

3.3 Emotion Label Representation

Two emotion label configurations are analyzed in this research project:

1. Fine-grained configuration, where all 28 emotion labels provided in the GoEmotions dataset are used.
2. Ekman-based configuration, wherein the emotion labels are collapsed into 7 emotion categories, according to the Ekman emotion theory [7], [21].

As depicted in **Table 2.A** below, semantically related fine-grained emotions are mapped to more general classes using the Ekman emotion categories. This approach mitigates label sparseness and promotes classification stability.

Some fine-grained emotions in the GoEmotions dataset lack an equivalent in Ekman's six basic emotions. The labels for such emotions are mapped to the most semantically related category.

Table 2.A Mapping from GoEmotions Labels to Ekman Emotion Categories

Ekman Category	Mapped GoEmotions Labels
Joy	admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride
Anger	anger, annoyance
Sadness	disappointment, grief, sadness, remorse
Fear	fear, nervousness
Disgust	disgust, disapproval
Surprise	realization, surprise, confusion, curiosity
Neutral	neutral

3.4 Model Architecture

For emotion classification using multiple labels, a RoBERTa architecture is utilized. The input text is tokenized and fed into the RoBERTa pre-trained model to get contextualized embeddings. The contextual embedding of the CLS

token is selected and passed through the dropout layer and the fully connected layer to get the logit for each emotion label.

To enhance generalization and decrease computational resources, the first six layers of the transformer model are frozen, whereas the rest of the layers are fine-tuned during the training phase.

Table 2 summarizes the model architecture.

Table 2. Model Configuration

Component	Description
Backbone	RoBERTa-base
Input length	128 tokens
Dropout	0.3
Output layer	Linear (7 labels)
Frozen layers	First 6 layers

3.5 Training Configuration

The training process employs Binary Cross-Entropy with Logits loss as the objective function since it is appropriate for multi-label classification problems. The model utilizes AdamW optimizer with a learning rate set to 2×10^{-5} .

Overfitting is mitigated through early stopping technique, where training is stopped if there is no improvement in the validation Micro-F1 score for three successive epochs.

Training parameters employed in the experiment are shown in **Table 3**.

Table 3. Training Parameters

Parameter	Value
Batch size	16
Learning rate	2×10^{-5}
Optimizer	AdamW
Loss function	BCEWithLogitsLoss
Epochs (max)	8
Early stopping	Yes
Patience	3
Device	GPU / CPU

The best model is selected based on validation Micro-F1 score.

The Micro-F1 score is used as the primary evaluation metric due to its suitability for imbalanced multi-label classification tasks.[21]

3.6 Threshold Optimization

In order to enhance the multi-label classification results, a threshold optimization technique is adopted for each emotion class individually after model training.

Instead of adopting a global fixed threshold, the best threshold for each emotion class is calculated based on the validation F1 score, and the thresholds can be varied from 0.1 to 0.9. The value of the threshold that provides the best F1 score will be adopted.

During validation, a default threshold of 0.3 is initially used, followed by label-wise threshold tuning to optimize performance for each class.

3.7 Training Behavior

In order to examine the convergence during the training process, the training loss is plotted per epoch, and the plot is illustrated in **Figure 3**.



Figure 3. Training Loss Curve

Training loss decreases gradually for each epoch in the above figure. This implies that the network is learning and making better predictions. The constant decrease in loss depicts stability in training.

4:Results and Discussion

The findings of the experiment as well as the effect of the representations of the emotions' labels and their thresholding optimization on the multi-label emotion classification task are discussed in this section.

4.1 Class Imbalance Investigation

The problem of strong class imbalance in the GoEmotions dataset is one of the most challenging problems to be solved during the multi-label emotion classification process.

As an approach to solving this problem, Ekman-based label reduction is employed. The class imbalance distribution before and after the procedure is depicted in **Table 4**.

Table 4. Class Imbalance Before and After Ekman Mapping

Setting	Min Samples	Max Samples	Imbalance Ratio
Original (Train)	1	14,218	14,218
Ekman (Train)	1,262	17,143	13.58
Original (Validation)	1	1,765	1,765
Ekman (Validation)	144	2,142	14.88
Original (Test)	1	1,787	1,787
Ekman (Test)	156	2,095	13.43

The results reveal a significant decrease in imbalance following the use of Ekman mapping, thus creating stability within the classification process.

4.2 Evaluation of Overall Model Performance

Overall model performance is displayed in Table 5 using the Micro-F1 and Macro-F1 scores. Table 5. Performance Comparison on GoEmotions Test Set

Model	Emotion Space	Micro-F1	Macro-F1
Hybrid (BERT+BiLSTM+Att)	28 emotions	0.5869	0.5269
RoBERTa	28 emotions	0.6055	0.5154
RoBERTa	7 emotions (Ekman)	0.6501	0.5951

RoBERTa performs better than the BERT+BiLSTM+Attention model in fine-grained classification, as shown in Table 5. Further gains are seen in applying Ekman-based emotion space mapping, especially in Macro-F1 scores.

4.3 Effect of Emotion Space Reduction

Classification performance can be enhanced by reducing the emotion space from 28 to 7, yielding more stable results.

The most prominent change is seen in an improvement in Macro-F1 scores, suggesting that there are fewer biases in class performance. In addition, the smaller semantic similarity between emotions facilitates better generalization.

4.4 Class-wise Performance Analysis

For a further understanding of model performance, we provide class-wise evaluation results for RoBERTa trained with Ekman-based emotion spaces in Table 6.

Table 6. Class-wise Performance (RoBERTa – Ekman Setting)

	Emotion	TP	FP	FN	TN	Precision	Recall	F1-score	Support
0	Joy	810	440	167	4008	0.6480	0.8291	0.7274	977
1	Sadness	223	281	475	4446	0.4425	0.3195	0.3710	698
2	Anger	1194	671	488	3072	0.6402	0.7099	0.6732	1682
3	Fear	485	264	165	4511	0.6475	0.7462	0.6934	650
4	Surprise	87	80	138	5120	0.5210	0.3867	0.4439	225
5	Disgust	152	108	120	5045	0.5846	0.5588	0.5714	272
6	Neutral	1400	922	365	2738	0.6029	0.7932	0.6851	1765

n from Table 6 below, the model shows relatively well-balanced results when dealing with emotions such as happiness, anger, and fear. Inadequate results when analyzing classes such as surprise and disgust can be explained by their low frequency in social networks.

4.5 Analysis of Confusion Matrices

For a better understanding of prediction results, confusion matrices are considered for each class of emotions separately, as presented in Figure 4.

Confusion Matrix for Joy:			Confusion Matrix for Anger:		
	Predicted	Predicted		Predicted	Predicted
	Negative	Positive		Negative	Positive
Actual			Actual		
Negative	4008	440	Negative	3072	671
Actual			Actual		
Positive	167	810	Positive	488	1194

Confusion Matrix for Sadness:			Confusion Matrix for Surprise:		
	Predicted	Predicted		Predicted	Predicted
	Negative	Positive		Negative	Positive
Actual			Actual		
Negative	4446	281	Negative	5120	80
Actual			Actual		
Positive	475	223	Positive	138	87

Confusion Matrix for Neutral:			Confusion Matrix for Fear:		
	Predicted	Predicted		Predicted	Predicted
	Negative	Positive		Negative	Positive

Actual Negative	2738	922	Actual Negative	4511	264
Actual Positive	365	1400	Actual Positive	165	485
Predicted Negative					
Predicted Positive					
Actual Negative					
Actual Positive					
Confusion Matrix for Disgust:					

Figure 4. Confusion matrices of the RoBERTa model for the Ekman-based emotion classification on the GoEmotions test set.

The confusion matrices of the RoBERTa model are further explored in order to analyze its behavior in each emotion class, as illustrated in Figure 3. The confusion matrix reveals information about true positives, false positives, false negatives, and true negatives for each emotion.

It is evident from Figure 3 that the model is highly efficient for emotion classes that have a high frequency of occurrence, such as joy, anger, and neutral. In these emotion classes, a high value of true positives is evident, indicating an efficient identification

process. However, moderate false positives are seen, indicating a certain degree of confusion between semantically similar emotions.

On the other hand, poor results are found in the categories of sadness, surprise, and disgust, where a high value of false negatives is seen, implying that the model fails to detect emotions properly. This can be due to their low frequency and subtle presence in social media texts..

4.6 Influence of Threshold Optimization

In order to achieve higher results, a label-based threshold optimization process is conducted.

Instead of applying the same global threshold to all emotions, a different threshold is optimized for each of them individually based on validation data. The goal of such a process is to ensure better balance between precision and recall of minor classes and contribute to improving Macro-F1 score overall.

5. Conclusion

The presented research examines how emotion label representation and threshold optimization affect multi-label emotion classification when using the GoEmotions dataset. It is evident that high class imbalance in fine-grained emotion scenarios negatively impacts classification results, especially in minority classes.

To deal with the problem of imbalance, the label reduction technique based on Ekman model was used, thus lowering the number of emotions from 28 to 7. Such an action helped balance the classes and made classification results much more consistent and reliable. Also, label-based threshold optimization helped improve the balance between precision and recall

The experiments showed that the highest accuracy was obtained by the RoBERTa model in the Ekman-based setting, with a Micro-F1 of 0.6501 and a Macro-F1 of 0.5951. It shows that the improvements related to labeling schemes and decision-making strategies might be more valuable than those in the model architecture.

Refreances:

- [1] S. A. Kumar S. and A. Geetha, "Emotion Detection from Text using Natural Language Processing and Neural Networks," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 14s, pp. 609–615, 2024.
- [2] D. Demszy, D. M. Jeongwoo, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions : A Dataset of Fine-Grained Emotions," pp. 4040–4054, 2020, doi: 10.18653/v1/2020.acl-main.372.
- [3] M. Kumar, "Emotion Recognition in Natural Language Processing: Understanding How AI Interprets the Emotional Tone of Text," *J. Artif. Intell. Cloud Comput.*, vol. 3, no. 6, pp. 1–5, 2024, doi: 10.47363/jaicc/2024(3)e238.
- [4] S. K. Bharti *et al.*, "Text-Based Emotion Recognition Using Deep Learning Approach," 2022, doi: 10.1155/2022/2645381.
- [5] L. R. Gomez *et al.*, "Emotion Recognition on Social Media Using Natural Language Processing (NLP) Techniques," *ACM Int. Conf. Proceeding Ser.*, no. DI, pp. 113–118, 2023, doi: 10.1145/3625156.3625173.
- [6] A. Thiab, L. Alawneh, and M. AL-Smadi, "Contextual emotion detection using ensemble deep learning," *Comput. Speech Lang.*, vol. 86, p. 101604, Jun. 2024, doi: 10.1016/J.CSL.2023.101604.
- [7] M. Rezapour, "E MOTION DETECTION WITH TRANSFORMERS : A," no. Xlm, 2024, doi: 10.48550/arXiv.2403.15454.
- [8] A. Bandhakavi, N. Wiratunga, and S. Massie, "Pattern Recognition Letters Lexicon based Feature Extraction for Emotion Text Classification," pp. 1–11, [Online]. Available: www.elsevier.com
- [9] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013, doi: 10.1111/j.1467-8640.2012.00460.x.
- [10] Y. Kim, "Convolutional Neural Networks for Sentence Classification," 2011.
- [11] R. Cascade-correlation and N. S. Chunking, "Neural Comput.-1997-LSTM-Long short term memory," vol. 9, no. 8, pp. 1–32, 1997.
- [12] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap.*, pp. 207–212, 2016, doi: 10.18653/v1/p16-2034.
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. doi: 10.48550/arXiv.1810.04805.
- [14] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [15] F. A. Acheampong, N.-M. Henry, and W. Chen, "Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches," 2021.
- [16] L. A. M. Bostan and R. Klinger, "An analysis of annotated corpora for emotion classification in text," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, pp. 2104–2119, 2018.
- [17] A. R. Murthy and K. M. Anil Kumar, "A Review of Different Approaches for Detecting Emotion from Text," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1110, no. 1, p. 12009, 2021, doi: 10.1088/1757-899x/1110/1/012009.
- [18] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.
- [19] A. Topic, M. Russo, M. Stella, and M. Saric, "Emotion Recognition Using a Reduced Set of EEG Channels Based on Holographic Feature Maps," *Sensors*, vol. 22, no. 9, 2022, doi: 10.3390/s22093248.
- [20] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning Based Text Classification: A Comprehensive Review," vol. 1, no. 1, pp. 1–42, 2021, [Online]. Available: <http://arxiv.org/abs/2004.03705>
- [21] F. Zhang, J. Chen, Q. Tang, and Y. Tian, "Evaluation of emotion classification schemes in social media text: an annotation-based approach," 2024. doi: 10.1186/s40359-024-02008-w.