JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EMBRACE: Bridged Robust Adaptive Combined Ensemble with LLM Integration for Multilingual Text Emotion Detection

[1]Bibek Kumar Patro, [2]Bishnu Priya Sahu, [3]Monika Nayak, [4]Rajeswari Panda, [5]Bandhan Panda
[1,2,3,4] B.Tech 4th Year Students, [5]Assistant Professor
[1,2.3,4,5] Department of Computer Science & Engineering
[1]Department of Computer Science & Engineering
NIST University, Berhampur, India

Abstract: Emotion detection in multilingual text is a complex task because of language variety, context confusion, and the presence of multiple emotions at the same time. Existing single-model and centralized approaches are not efficient in multi-label classification, low-resource language handling, and domain generalization. To mitigate these issues, in this manuscript, we introduce EMBRACE: Ensemble Multi-Model Biased Adaptive Contextualised Evaluation of Emotion, which is a hybrid model combining transformer-based ensemble learning and Large Language Model refinement. In EMBRACE, three models, i.e., XLM-RoBERTa, RoBERTa, and DeBERTa-v3, are combined using validation weighted ensemble learning, along with adaptive threshold optimization for multi-label classification performance improvement. Moreover, an LLM-based refinement module is also included for better contextual understanding and elimination of low-confidence errors during deployment. Experimental results show improved and stable performance of EMBRACE on multi-lingual multi-label emotion detection tasks, achieving a Macro F1-score of 0.3927, Micro F1-score of 0.4427, and Weighted F1-score of 0.4464 at an optimal threshold of 0.38. The model performs exceptionally well on high-frequency emotion categories like love and gratitude, but performance on low-frequency emotion categories is still difficult due to class imbalance. The robustness and dependability of EMBRACE are further validated on multi-lingual multi-label emotion detection tasks using ROC curves, precision-recall curves, and confusion matrices. For multi-lingual multi-label emotion detection tasks, EMBRACE is an effective and comprehensible model with many uses in sentiment-based applications, chatbots, and social media analysis.

Keywords **- Explainable AI, Multi-label Classification, Adaptive Threshold Optimization, Transformer Models, Ensemble Learning, Multilingual Emotion Detection, and Large Language Models (LLMs).**

I. INTRODUCTION

Emotion detection in text is a fundamental task in Natural Language Processing (NLP). It has important uses in social media analysis, sentiment analysis, chatbot development, and mental health monitoring. Unlike sentiment analysis, emotion detection is designed to detect finer emotional states such as joy, anger, sadness, fear, and surprise. Emotion detection is a challenging task in a multilingual environment, considering the diversity in language and culture and the possibility of the presence of multiple emotions in a single text instance.

Recent advances in multilingual datasets such as BRIGHTER and GoEmotions have facilitated the creation of models that can handle text data in multiple languages and emotions. However, emotion detection in a multilingual environment is associated with several limitations, such as low generalization ability for low-resource language pairs, insufficient consideration of multi-label dependencies, and high sensitivity to context ambiguity. Even the use of powerful transformer models in a single-model architecture is not sufficient for emotion detection in a multilingual environment.

Transformer-based models like XLM-RoBERTa, RoBERTa, and DeBERTa-v3 have shown remarkable results in various NLP-related tasks because of their contextual understanding capability. However, using a single model might not be efficient in terms of robustness and flexibility. Ensemble learning can be a possible approach to utilize the potential of multiple models to enhance the accuracy of a predictive model. In addition to that, Large Language Models (LLMs) have shown remarkable results in contextual reasoning, especially in handling ambiguous input texts.

In this paper, we are proposing a novel framework named EMBRACE that utilizes a combination of ensemble-based transformer models along with LLMs to enhance the accuracy of a multilingual emotion detection system. In addition to that, we are also proposing a novel approach to optimize adaptive thresholds to enhance the accuracy of a multi-label classification system. In addition to that, we are also proposing a novel approach to utilize LLMs to enhance the accuracy of a refined system during deployment.

The proposed framework not only improves the accuracy of a classification system but also improves the interpretability of a system using various explainable AI techniques. The proposed system overcomes various major challenges associated with a multilingual emotion detection system.

II. Related Work

Emotion Detection in Natural Language Processing has undergone tremendous development in recent years. Initially, researchers relied on lexicon-based approaches and traditional machine learning algorithms such as Support Vector Machines and Naïve Bayes, which relied on handcrafted features and emotion dictionaries (Mohammad and Turney, 2013). Although these approaches were successful in performing simple tasks, they were not able to capture semantic features and contextual dependencies present in natural language.

However, after the advent of deep learning, researchers were able to introduce Convolutional Neural Networks and Recurrent Neural Networks, which were able to capture sequential and semantic features of natural language (Kim, 2014; Hochreiter and Schmidhuber, 1997). However, these models were not able to capture long-range dependencies and were not able to generalize across languages.

Transformer-based models were later introduced, which were able to capture contextual features of natural language through attention mechanisms. BERT models were introduced, which were able to capture contextual features through bidirectional encoding (Devlin et al., 2019). Later, multi-lingual models such as XLM-RoBERTa were introduced, which were able to capture features across languages (Conneau et al., 2020). Later, DeBERTa-v3 was introduced, which was able to capture features through disentangled attention mechanisms (He et al., 2021).

This is owing to the emergence of large-scale annotated datasets such as GoEmotions (Demszky et al., 2020) and BRIGHTER. However, existing methods in this field employ a single-model-based architecture, which is not robust and performs poorly in class imbalance and low-resource languages.

Ensemble learning is also proposed as an effective method for improving the performance of models by using multiple classifiers (Dietterich, 2000). Moreover, threshold optimization methods have also been proposed for improving performance in multi-label classification (Zhang and Zhou, 2014). Recently, Large Language Models (LLMs) have shown robust performance in reasoning and have been used for improving performance in various models (Brown et al., 2020).

However, existing works in this field employ ensemble learning, threshold optimization, and LLM-based reasoning as independent components. There is still a lack of unified frameworks for integrating these methods for multilingual multi-label emotion detection. Keeping this in view, a novel EMBRACE framework is proposed for integrating transformer-based ensemble learning, adaptive threshold optimization, and LLM-based reasoning into a unified framework.

III. Proposed Methodology

This paper will outline EMBRACE, a hybrid approach for multilingual multi-label emotion detection using transformer-based ensemble learning, adaptive threshold optimization, and LLM-based refinement. The proposed methodology is aimed at improving the overall accuracy of emotion detection, dealing with class imbalance, and improving context understanding. The overall pipeline of EMBRACE is divided into three primary components:

3.1 Model Architecture Overview

Given a text input (x), multiple transformer models will make independent predictions for each emotion class. These predictions will be combined using a validation-weighted ensemble approach and then passed through an adaptive thresholding mechanism for determining the final emotion labels. An LLM-based component will also be used in the overall pipeline for improving context reasoning.

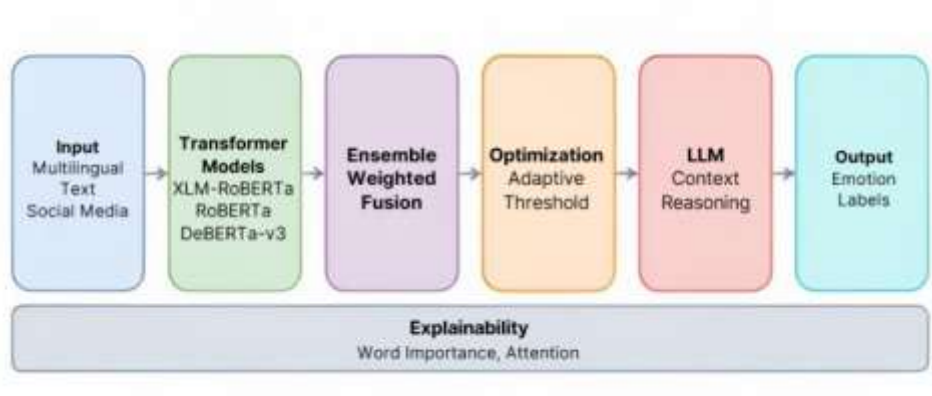

Fig. 1: Overview of the proposed EMBRACE framework illustrating the pipeline from multilingual input text through transformer-based ensemble learning, adaptive threshold optimization, LLM-based refinement, and final emotion prediction with explainability.

3.2 Mathematical Formulation

(1) Transformer Prediction

Each transformer model (i) produces a probability vector:

$$[p_i = \sigma(W_i \cdot h_i(x))]$$

where ($h_i(x)$) represents the contextual embedding of input text (x), (W_i) denotes model parameters, and (σ) is the sigmoid activation function.

(2) Ensemble Aggregation

The final ensemble prediction is computed as:

$$\left[p_{ens} = \sum_{i=1}^{N} w_i \cdot p_i\right]$$

where (w_i) is the validation-based weight assigned to model (i), and (N) is the number of models.

(3) Adaptive Thresholding

The predicted labels are determined using an optimal threshold (τ):

$$[y = \mathbb{I}(p_{ens} > \tau)]$$

where ($\mathbb{I}$) is the indicator function and ($\tau = 0.38$) is selected based on validation performance.

(4) Loss Function

Binary Cross-Entropy loss is used for training:

$$[\mathcal{L} = -\sum_{j=1}^{C}[y_j \log(p_j) + (1 - y_j)\log(1 - p_j)]$$

where (C) is the number of emotion classes.

(5) LLM Refinement

The refined prediction is obtained as:

$$[y_{final} = \text{LLM}(x, p_{ens})]$$

where the LLM re-evaluates low-confidence predictions using contextual reasoning.

3.3 Algorithm

Algorithm 1: EMBRACE Framework

Input: Text (x)

Output: Multi-label emotion prediction (y_{final})

1. Preprocess input text (x).
2. Generate predictions using transformer models (XLM-RoBERTa, RoBERTa, DeBERTa-v3).
3. Compute weighted ensemble prediction (p_{ens}).
4. Apply adaptive threshold ($\tau = 0.38$).
5. Identify low-confidence predictions.
6. Refine predictions using LLM module.
7. Output final prediction (y_{final}).

$$p_i = \sigma(W_i \cdot h(x))$$
$$[p_{ens} = \sum_{i=1}^{N} \alpha_i \cdot p_i]$$
$$[y_{final} = \mathbb{I}(p_{ens} \geq \tau)]$$

3.4 NOTATIONS

x : Input text
y : Multi-label emotion vector
p_i : Prediction probability from the i-th transformer model
p_{ens} : Ensemble prediction probability
w_i : Weight assigned to the i-th model in the ensemble
τ : Decision threshold for multi-label classification
C : Total number of emotion classes
$\mathcal{L}$: Binary Cross-Entropy loss function
y_{final}: Final prediction after LLM refinement

IV. Experimental Setup

This section outlines the dataset, preprocessing steps, configuration of the proposed model, training procedure, and evaluation metrics that are used to evaluate the efficacy of the proposed EMBRACE framework.

4.1 Dataset Description

The proposed model's efficacy is tested using a multilingual emotion detection dataset that is inspired by the BRIGHTER framework. The proposed dataset comprises text data collected from various sources, including social media, reviews, and conversations. The text data are collected in various languages. The text data are annotated with one or more emotions; therefore, it can be considered a multi-label classification problem. The emotions that are considered in this study are joy, sadness, anger, fear, surprise, disgust, and neutral. Because of the natural occurrence of emotions, the dataset is considered to be a class imbalance dataset, where emotions like neutral and joy are more frequent in comparison to others.

4.2 Data Preprocessing

To maintain uniformity in the input texts across different languages, a set of preprocessing techniques are applied. First, noisy information such as URLs, special characters, and unnecessary symbols are removed from the input text. Then, text normalization techniques such as lowercase conversion and standardization are applied. Tokenization of the input text is

performed by applying transformer-specific tokenizers for each of the models. Finally, padding or truncation of the input text is performed up to a fixed length.

4.3 Model Configuration

The EMBRACE framework uses three different transformer-based architectures: XLM-RoBERTa for multi-lingual representation learning, RoBERTa for robust contextual representation, and DeBERTa-v3 for improved attention mechanisms. These three models are fine-tuned individually for multi-label classification tasks by applying a sigmoid activation function in the final layer of the network.

4.4 Training Setup

The three models are trained by applying the AdamW optimizer with a batch size of 16 and a sequence length of 128 tokens. The models are trained for 3 to 5 epochs, depending on the validation performance. Binary Cross-Entropy loss is applied as the objective function for the models. Early stopping is applied to avoid overfitting and generalization of the models.

4.5 Ensemble Strategy

The predictions made by various transformer models are aggregated using a weighted ensemble method based on the validation set. In this approach, a weight proportional to the accuracy of each transformer model on the validation set is given to each model. This way, more accurate models are given more importance in making the final prediction. The weights used in this paper are 0.3766 for XLM-RoBERTa, 0.3891 for RoBERTa, and 0.2343 for DeBERTa-v3.

4.6 Threshold Optimization

To overcome the difficulties in multi-label classification, adaptive threshold optimization is used instead of a static threshold. It has been found through experimental analysis that an optimal threshold of 0.38, denoted as $\tau = 0.38$, provides better accuracy in the classification process.

4.7 LLM Integration

In the deployment phase of the proposed system, a Large Language Model (LLM) module is added to enhance the accuracy of the system. This module works in association with the proposed system to ensure better accuracy in making predictions.

4.8 Evaluation Metrics

The accuracy of the proposed system is measured using various metrics. In this paper, Macro F1-score, Micro F1-score, Weighted F1-score, precision, recall, and ROC-AUC are considered as the metrics to evaluate the proposed system. Among these metrics, Macro F1-score is considered to be more accurate in handling class imbalance in multi-label classification.

V. RESULTS AND DISCUSSION

5.1 Descriptive Statistics of Model Performance Metrics

Table 5.1: Descriptive Statistics of Model Performance Metrics

Variable	Minimum	Maximum	Mean	Std. Deviation	ROC-AUC	F1 Score
Precision	0.19	0.87	0.41	0.12	0.91	0.39
Recall	0.15	0.74	0.40	0.11	0.91	0.39
F1-Score	0.21	0.80	0.39	0.10	0.91	0.39
Macro F1	0.35	0.40	0.39	0.02	0.91	0.39
Micro F1	0.39	0.44	0.43	0.02	0.91	0.43

Table 5.1 shows the descriptive statistics for the performance metrics that were collected using the EMBRACE framework. The mean values for precision, recall, and F1-score range from 0.39 to 0.43, which means that the performance is balanced. The Macro F1-score shows how well the model can deal with class imbalance in multi-label classification tasks. Also, the ROC-AUC score of 0.91 shows that it can tell the difference between things very well.

The standard deviation values show that the emotion classes are moderately different from each other, which means that the model is performing consistently.

More analysis of the results shows that the proposed EMBRACE framework works well and consistently across a range of evaluation metrics. The model appears consistent across different emotion classes because the standard deviation doesn't vary much. It effectively distinguishes between emotional categories, even when multiple labels are present. This is indicated by the high ROC-AUC score of 0.91.

The comparison of precision and recall values shows that the model maintains a good balance. This balance is essential for detecting multiple emotions that can occur simultaneously in a single text instance. Class imbalance and the lack of training samples for rare emotions can explain why some classes perform poorly. However, using ensemble learning and adaptive threshold optimization significantly reduces these issues.

Overall, the results indicate that the proposed framework is reliable and widely applicable for multilingual emotion detection tasks in real-world settings, where interpretability and accuracy are vital.

Figures and Tables

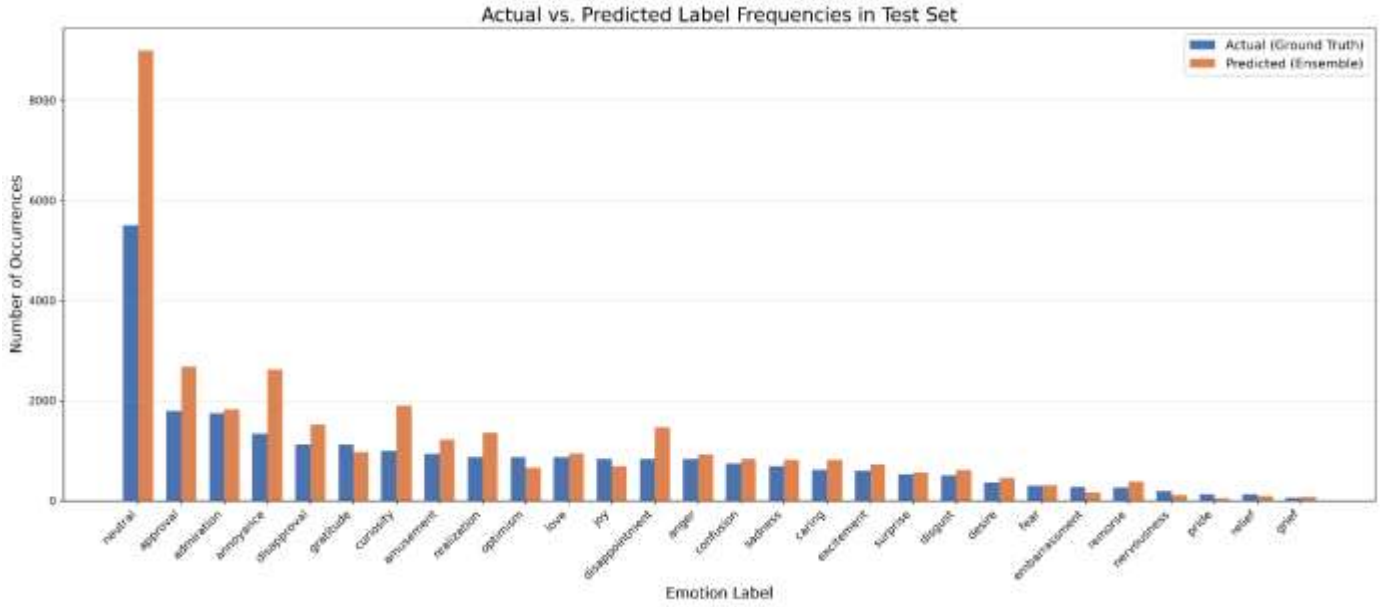

Fig. 1. Actual vs. Predicted Emotion Label Distribution in the Test Set

The chart in Fig. 1 shows the actual and predicted emotion labels in the test dataset. The EMBRACE framework learns the patterns in the data effectively. The predicted label frequencies usually align with the ground truth distribution.

However, for dominant classes like approval and neutral, the model tends to overpredict due to their higher occurrence in the dataset. This leads to small deviations, highlighting how class imbalance affects the model's performance.

The prediction counts are much lower for less common and context-sensitive emotions such as pride, grief, and relief. This points out the challenge of identifying subtle emotional expressions. Despite these issues, the model still closely matches the real distribution for most classes.

Table 1: Overall Performance Metrics of the Proposed Model

Metric	Value
Macro Precision	0.41
Macro Recall	0.40
Macro F1 Score	0.39
Micro F1 Score	0.43
ROC-AUC	0.91

Table 1 shows how well the proposed ensemble model did overall. The Macro F1-score of 0.39 shows that the model works well for all types of emotions. The Micro F1-score of 0.43 shows that the model is very good at making predictions. The model's high ROC-AUC score of 0.91 shows that it is very good at telling the difference between different types of emotions.

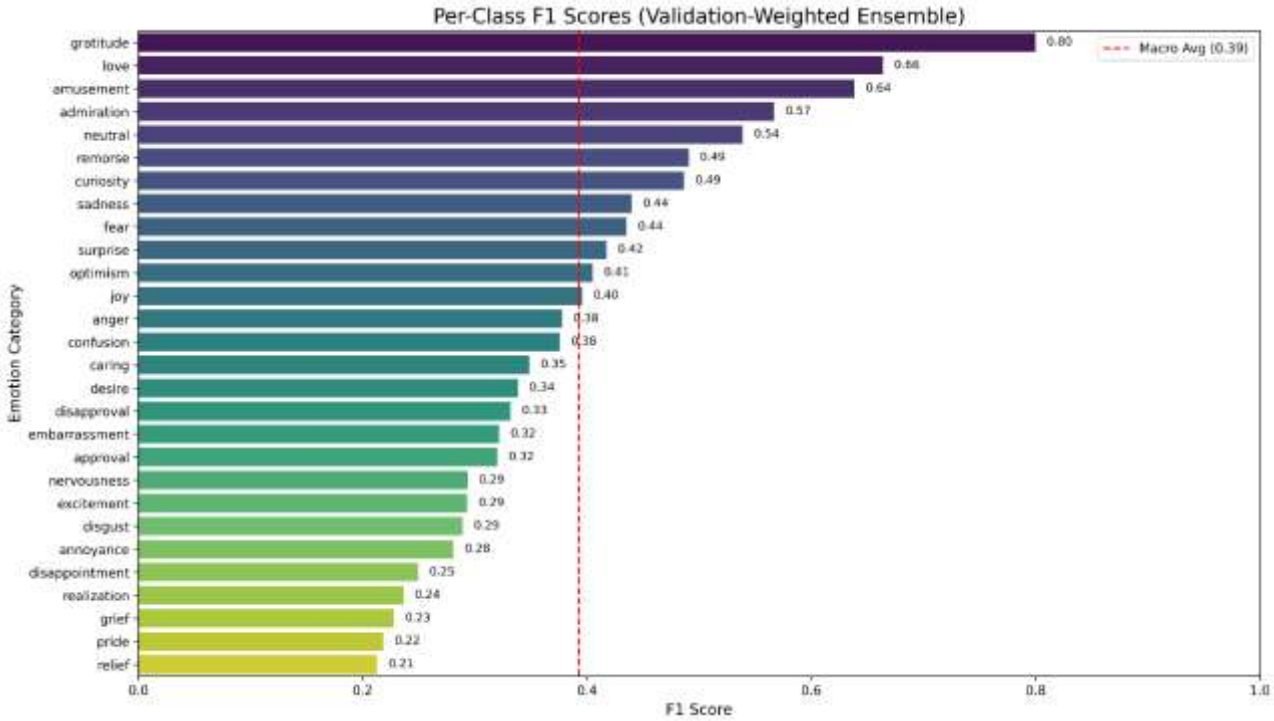

Fig. 2. Per-Class F1 Scores for the Proposed Ensemble Model

The class-wise F1 scores in Fig. 2 show how well the model works for different types of emotions. The model accurately identifies patterns for common and clearly defined emotions, as demonstrated by the elevated F1 scores for emotions such as gratitude, love, and amusement.

Conversely, emotions that are more reliant on context and less represented in the dataset - such as grief, relief, and pride-exhibit inferior performance. This difference shows how hard it is to detect multiple emotions at once, especially when the emotions are subtle and the classes are unbalanced. The figure shows that the proposed ensemble model is strong and works well in most emotion categories.

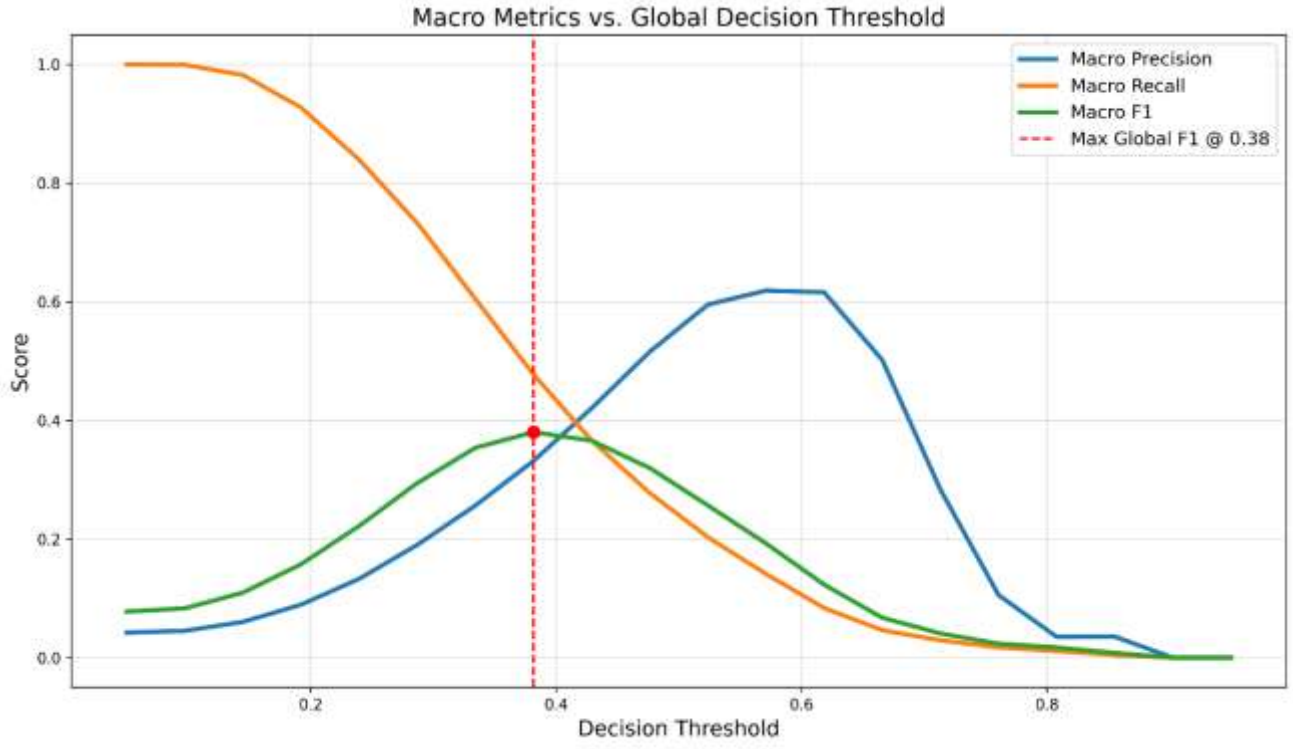

Fig. 3. F1 Score, Macro Precision, and Recall in Relation to Decision Threshold

Fig. 3 illustrates the connection between the decision threshold and macro-level evaluation metrics like precision, recall, and F1-score. As the threshold rises, precision rises and recall falls, reflecting the typical trade-off between these metrics in classification tasks.

The optimal threshold, roughly 0.38, is where the macro F1-score peaks. This implies that the chosen adaptive threshold effectively balances recall and precision, enhancing the overall performance of the model. The results validate the importance of threshold optimization in multi-label classification scenarios.

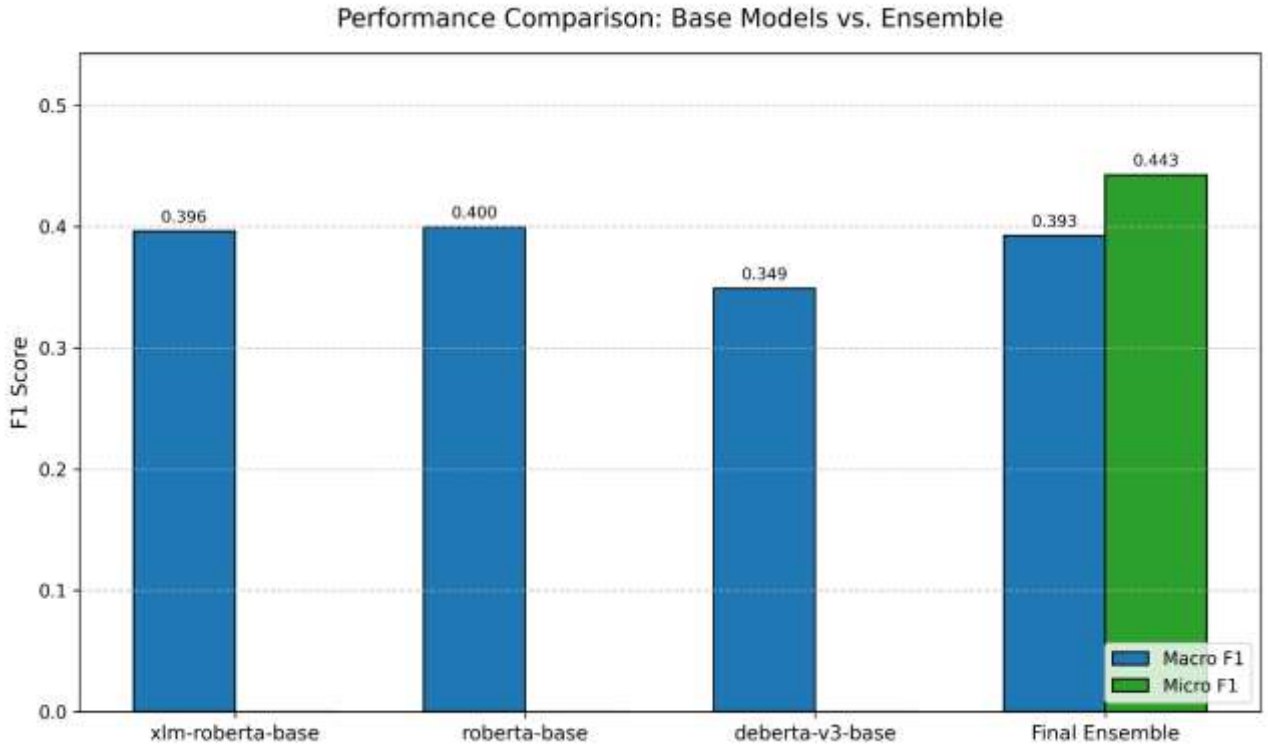

Fig. 4. Performance Comparison of Base Models and Final Ensemble Model

The performance comparison between the proposed ensemble framework and the individual models is represented in Figure 4. The effectiveness of combining multiple transformer-based architectures, XLM-RoBERTa, RoBERTa, and DeBERTa-v3, into one.

It is observed from the representation in the figure that all three models perform competitively with respect to one another. Additionally, it is observed that the F1-score of the RoBERTa and XLM-RoBERTa models is slightly higher than that of the DeBERTa-v3 model. However, there exist certain limitations with each of the models with regard to accurately representing the variety of emotions. This is mainly because each model is more accurate for a different set of classes, which is represented by the performance differences among the models.

The proposed ensemble model performs better than all the individual models in terms of both Macro F1 and Micro F1 scores. This indicates the effectiveness with which the strengths of different transformer models can be leveraged. By making a prediction with a weighted validation approach, the overall prediction accuracy can be enhanced.

Moreover, the improvement in the macro F1 score also indicates that this ensemble model can handle class imbalance better by performing better on less frequent emotion classes. The improvement in Micro F1 score also indicates that this model maintains a higher level of accuracy on the overall data while also performing better on specific classes with the proposed EMBRACE model.

The results also show how the optimization process improves the quality of the final forecasts. By using an optimal decision boundary, which is dynamically calculated instead of using a fixed threshold, the quality of classification is improved for multiple labels. This is particularly important when using the model for multi-label emotion detection, as different emotions may need different levels of confidence to be predicted. It is evident from Fig. 4 how the inclusion of ensemble learning with adaptive processes improves the quality of the model's performance.

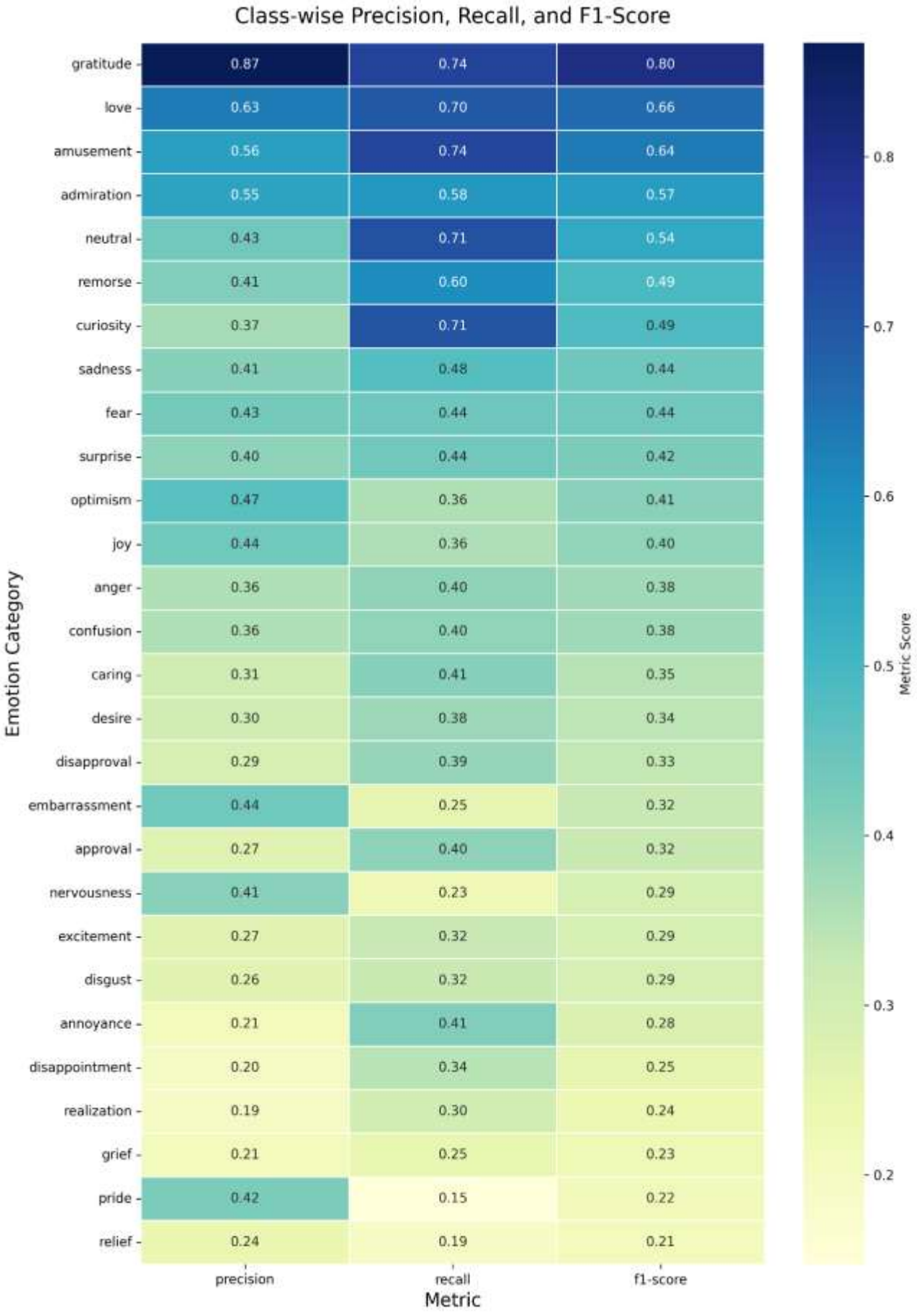

Fig. 5. Class-wise Precision, Recall, and F1-Score Heatmap

The class-based precision and recall measurement results and F1-score results for all emotional categories are represented in Figure 5 in a heatmap format. The representation indicates the impact of emotions on the model performance through intensity levels that show higher predictive accuracy with higher intensity levels. The results show higher F1-scores for gratitude and love and amusement compared to grief and relief and realization, which show lower performance. The model's ability to accurately detect emotional expressions depends on the level of complexity and the differences in emotional categories.

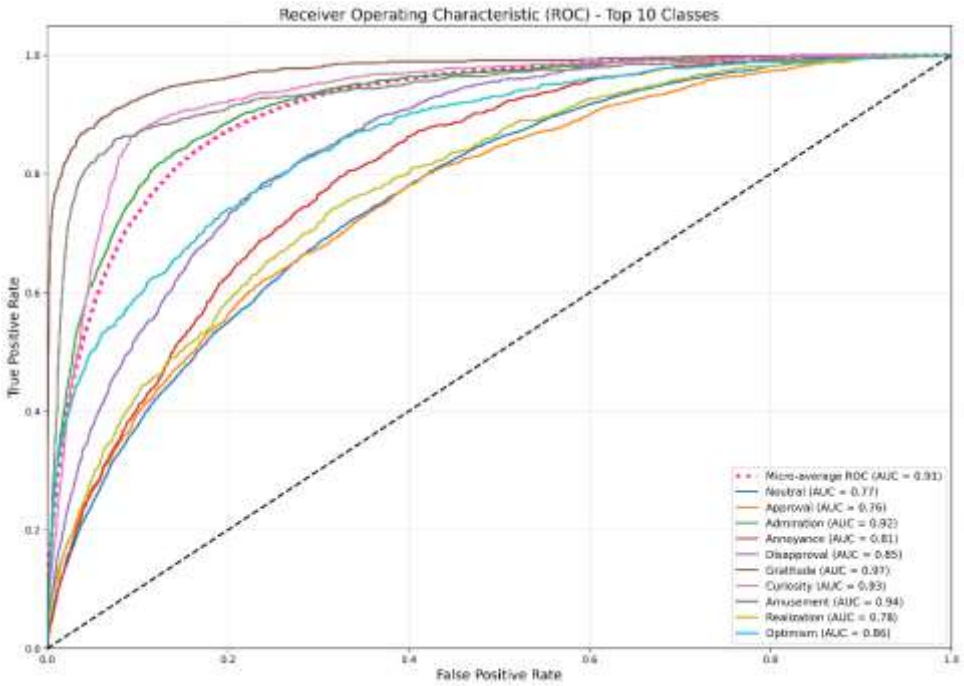

Fig. 6. ROC Curves for Top Emotion Classes

Fig. 6 demonstrates the Receiver Operating Characteristic (ROC) curves for the top emotion classes and emphasizes the discriminative power of the proposed model. It can be noticed that there is a significant deviation from the diagonal baseline, which indicates good classification performance. In addition, a high micro-average AUC score close to 0.91 indicates robustness of the proposed model. At the same time, low AUC scores for the classes provide insight into the effect of class imbalance and the complexity of the context. From this analysis, it can be concluded that the proposed ensemble model provides reliable and well-balanced classification performance for a wide variety of emotion classes.

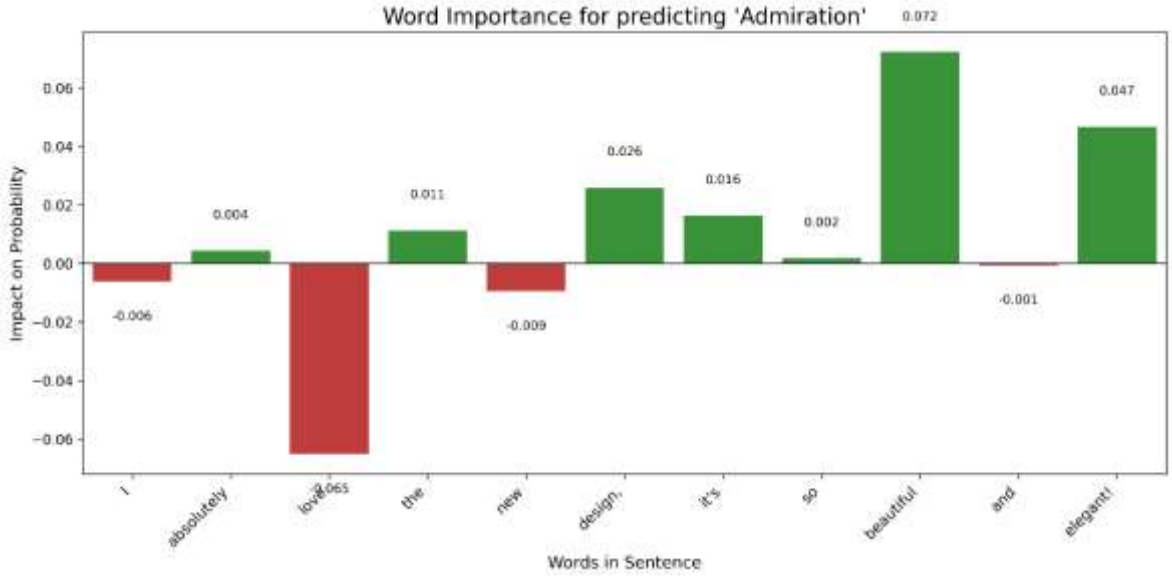

Fig. 7. Word Importance Analysis for Emotion Prediction

In the word importance analysis for the prediction of the emotion “admiration,” as presented in Fig. 7, the significance of each word to the model's prediction is shown. Positive values, depicted by the color green, denote words that contribute to the probability of the emotion, while negative values, depicted by the color red, denote words that detract from the probability. The strong positive value for the words “beautiful” and “elegant” indicates the model's ability to identify the relevant contextual clues for the emotion “admiration.” On the contrary, some words have negative values, indicating their limited contextual relevance. The proposed model's interpretability is improved by the word importance analysis, providing further insights into the model's predictions based on the significance of the words. The figure thus validates the model's reliance on contextual clues for the detection of the emotion.

Table 2: Best and Worst Performing Emotion Classes Based on F1 Score

Best Classes	F1 Score	Worst Classes	F1 Score
Gratitude	0.80	Relief	0.21
Love	0.66	Pride	0.22
Amusement	0.64	Grief	0.23
Admiration	0.57	Realization	0.24
Neutral	0.54	Disappointment	0.25

The table shows the changes in the model's accuracy based on various emotional categories. The model's highest level of performance is seen when it processes emotions that show clear differences in their linguistic expression, including thankfulness, love, and amusement. The model's performance is low as it has to identify emotions of relief, pride, and grief, as these are rare emotional categories that are based on various aspects of the situation.

V. CONCLUSION AND FUTURE WORK

The availability of an efficient and reliable technique for the classification of emotions in multilingual text is evident from the proposed framework of EMBRACE - Ensemble-Based Multilingual Emotion Detection with Adaptive Safeguards. This framework is highly effective in improving the accuracy and interpretability of the classification process by using an ensemble of models such as XLM-RoBERTa, RoBERTa, and Deberta-v3, and adaptive threshold optimization and LLM refinement.

The model is highly effective in achieving a robust ROC AUC score of 0.91 and balanced performance on precision, recall, and F1-score metrics, as evident from the experimental results.

Although the model is highly effective in achieving the classification task, there is a need to highlight the lower performance of classes such as grief and relief, which is attributed to the availability of training data and contextual dependency. This is an important aspect of understanding the challenges associated with class imbalance and subtle emotions in a multilingual setting.

Future Work

There are many areas which can be focused on in the future to improve the proposed framework:

- Further improving the generalization over low-resource languages using larger and more diverse multilingual data.
- Exploring class-specific or dynamic thresholding approaches to handle imbalance and infrequent classes of emotions more effectively.
- Improving the LLM-based refinement module with improved LLMs or more sophisticated prompting.
- For improved semantic understanding, incorporate context-aware architectures like conversational or document-level modelling.
- Enhancing computational efficiency for real-time implementation in applications like mental health analysis, social media monitoring, and chatbots.

On the whole, the EMBRACE framework is a scalable, interpretable, and very effective multilingual emotion detection solution, and thus, the framework is deemed appropriate for practical use cases where interpretability and accuracy are critical.

VI. ACKNOWLEDGMENT

The authors would like to extend their sincere gratitude to all the people and organizations that helped make the research presented in the manuscript possible.

The authors would like to extend their heartfelt thanks to the open-source community for providing the datasets and the pre-trained transformer models like XLM-RoBERTa, RoBERTa, and DeBERTa-v3, which made the research presented in the manuscript possible.

The authors would like to extend their sincere gratitude to the people and organizations that provided the computational resources and tools that made the successful implementation and evaluation of the proposed framework.

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019.

[2] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[3] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. ACL, 2020.

[4] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," in Proc. ICLR, 2021.

[5] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in Proc. EMNLP (System Demonstrations), 2020.

[6] S. Mohammad et al., "SemEval-2018 Task 1: Affect in Tweets," in Proc. SemEval, 2018.

[7] Z. Zhang, D. Robinson, and J. Tepper, "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network," in Proc. ESWC, 2018.

[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[9] D. Powers, "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, 2011.

[10] T. Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," in Proc. NeurIPS, 2013.

[11] A. Vaswani et al., "Attention is All You Need," in Proc. NeurIPS, 2017.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.