A Comparative Study of Deep Learning Models and Data Augmentation for Imbalanced Multi-Label Emotion Classification

Herlina
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
herlina.lim@ugm.ac.id

Findra Kartika Sari Dewi
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
findrakartikasaridewi@ugm.ac.id

Marwan Ramdhany Edy
Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia
marwanramdhanyedy@ugm.ac.id

***Abstract*—Multi-label emotion classification identifies multiple co-occurring emotional states in text, often complicated by class imbalance. This study evaluates TextCNN and efficient Transformer-based models—DistilBERT, DistilRoBERTa, and ALBERT—across seven emotion categories: Admiration, Amusement, Gratitude, Love, Pride, Relief, and Remorse. Data augmentation is applied to address imbalance, and performance is assessed using standard multi-label metrics and confusion matrices. Results show Transformer-based models outperform TextCNN, with DistilBERT achieving the highest Micro-F1 and consistent per-label performance. Augmentation improves recall for minority emotions such as Gratitude, Love, and Remorse but may increase false positives, revealing a trade-off between sensitivity and precision. These findings highlight the importance of efficient Transformer architectures combined with carefully designed augmentation strategies for robust and balanced multi-label emotion classification.**

Keywords—multi-label emotion classification, natural language processing, class imbalance, data augmentation, Transformer models

I. Introduction

Classification of emotions in text is a critical task in Natural Language Processing (NLP), facilitating a broad range of downstream applications such as emotional chatbots, stock market prediction, and large-scale social media monitoring [1], [2], [3]. By enabling machines to recognize and interpret affective signals expressed in language, emotion classification supports more human-centered and context-aware intelligent systems. While early approaches often treated emotion recognition as a single-class problem, human communication frequently conveys multiple affective states simultaneously. Consequently, modern research predominantly focuses on Multi-Label Classification (MLC), where a single text instance can be mapped to a subset of co-existing emotion labels [3], [4], [5].

Multi-label emotion classification presents unique challenges compared to traditional single-label tasks. Emotions are inherently subjective, overlapping, and context-dependent, making their automatic detection particularly difficult. A short text may simultaneously express admiration and gratitude, or relief and remorse, requiring models to capture nuanced semantic and emotional cues beyond surface-level lexical patterns. These challenges are further amplified in real-world datasets, where emotional expressions are often sparse, informal, and highly diverse [5], [6].

This project centers on the MLC task as defined in the Shared Task of Emotion Recognition (INFO 557 Fall 2025), hosted on the Codabench platform. The task aims to classify input text into seven fine-grained emotion categories, namely admiration, amusement, gratitude, love, pride, relief, and remorse. These emotion labels reflect subtle affective distinctions and frequently co-occur within a single textual instance, making the task a suitable benchmark for evaluating the effectiveness of modern multi-label emotion classification models.

To address the complexity of this task, we evaluate several state-of-the-art deep learning architectures that have demonstrated strong performance in text classification and emotion recognition. In particular, Transformer-based models have become the dominant paradigm in NLP due to their ability to capture contextualized semantic representations through self-attention mechanisms [6], [7]. However, their large model size and computational requirements pose practical challenges. Therefore, this study focuses on efficient Transformer variants, including DistilBERT, DistilRoBERTa, and ALBERT, which aim to balance performance and computational efficiency. In addition to Transformer-based approaches, we also employ TextCNN as a classical convolutional baseline, given its proven effectiveness in extracting salient local patterns from text [6].

Despite the strong representational capabilities of deep neural models, preliminary analysis indicates potential performance limitations in predicting certain minority emotion classes, particularly Pride and Relief. This issue may be attributed to class imbalance, a common characteristic of emotion datasets in which a small number of labels dominate the training distribution, while others appear infrequently[8]. Such long-tail distributions often bias learning algorithms toward majority classes, leading to reduced generalization performance for underrepresented emotions [9]. Addressing data imbalance is therefore an important aspect of building robust multi-label emotion classifiers. Prior studies have shown that techniques such as data augmentation and loss function optimization can help mitigate the negative effects of imbalanced label distributions by improving model exposure to minority classes and encouraging more balanced learning dynamics [1], [8]. Motivated by these findings, this project explores targeted strategies to enhance the model's ability to learn from underrepresented emotion categories, with the goal of improving both per-label performance and overall classification robustness.

Overall, this work aims to provide an empirical evaluation of efficient deep learning models for multi-label emotion classification under realistic dataset constraints. By systematically comparing different neural architectures and

investigating strategies to address class imbalance, this study contributes practical insights into the design of effective and computationally efficient emotion recognition systems.

II. Literature Review

A. Multi-Label Classification Frameworks

Multi-label classification (MLC) is a fundamental framework for emotion analysis, where a single text instance may simultaneously express multiple emotional states[10]. Unlike single-label classification, which assumes mutual exclusivity among classes, MLC allows labels to overlap, making it particularly suitable for modeling complex affective expressions in natural language. In emotion datasets, it is common for emotions such as admiration, gratitude, and love to co-occur within the same textual context, reflecting the multifaceted nature of human emotional expression [10], [11].

In neural-based MLC, the problem is commonly addressed using a unified model architecture with independent sigmoid outputs for each label, enabling the model to predict multiple emotions concurrently [11], [12]. This formulation treats each label as an independent binary classification task while sharing a common feature representation across all labels. Compared to traditional problem transformation techniques, such as binary relevance or classifier chains, this approach offers improved scalability and simpler implementation in deep learning settings [6], [11]. As a result, it has become the dominant paradigm for multi-label emotion classification in recent studies.

Beyond standard sigmoid-based formulations, recent studies have also explored label-aware architectures that incorporate explicit label representations to enhance multi-label text classification performance [13]. These approaches aim to leverage semantic relationships among labels to improve prediction quality, particularly in scenarios where label dependencies are strong. Although such methods are not universally adopted, they highlight the ongoing research interest in improving representation learning for complex MLC tasks.

B. Neural Architectures for Text-Based Emotion Classification

Deep neural networks have substantially advanced emotion classification by enabling automatic representation learning directly from raw text, thereby reducing reliance on handcrafted linguistic features and task-specific heuristics[6], [10]. By learning hierarchical and distributed representations, these models are better equipped to capture subtle emotional cues, contextual dependencies, and semantic variations present in natural language.

1) Convolutional Neural Networks

Convolutional Neural Networks (CNNs), particularly TextCNN, are among the earliest deep learning architectures successfully applied to emotion classification tasks[6]. TextCNN employs multiple convolution filters with varying kernel sizes to capture salient local patterns, such as emotionally informative n-grams and short phrase structures. These features are subsequently aggregated using max-pooling operations to form a fixed-length sentence representation.

Despite its architectural simplicity, TextCNN remains a competitive baseline in emotion classification due to its computational efficiency, robustness, and relatively low training cost [6]. Its ability to effectively model local textual patterns makes it especially suitable for short texts, which are commonly encountered in emotion recognition datasets.

2) Transformer-Based Models and Transfer Learning

Transformer-based architectures have become the dominant paradigm in natural language processing through their reliance on self-attention mechanisms, which enable global context modeling without recurrent structures [1], [6]. Models such as BERT are pre-trained on large-scale unlabeled corpora, allowing them to learn rich, contextualized word representations that capture both syntactic and semantic information. These representations can then be fine-tuned for downstream tasks, including emotion classification [6], [12], [14].

Transfer learning plays a crucial role in the success of Transformer-based models, as it significantly reduces the dependence on large task-specific labeled datasets. This is particularly beneficial for emotion classification tasks, where labeled data may be limited or unevenly distributed across emotion categories [12], [14].

3) Efficient Transformer Variants

Despite their strong performance, full-sized Transformer models are computationally expensive in terms of memory usage and inference time. To address these limitations, several efficient Transformer variants have been proposed. DistilBERT utilizes knowledge distillation techniques to compress the original BERT model while retaining most of its representational capacity [5]. DistilRoBERTa applies similar distillation strategies to the RoBERTa architecture, offering improved efficiency with minimal performance degradation[7]. ALBERT further reduces computational overhead through parameter sharing and factorized embedding representations[7].

These lightweight Transformer models are particularly well-suited for large-scale or resource-constrained multi-label emotion classification settings, where efficiency, scalability, and faster inference are important practical considerations.

C. Data Imbalance in Multi-Label Emotion Classification

A major challenge in multi-label emotion datasets is label imbalance, where certain emotions occur significantly less frequently than others [8], [9]. This long-tail distribution often leads to biased learning behavior, causing models to favor dominant labels while underperforming on minority emotion classes[14]. As a result, overall evaluation metrics may mask poor performance on less frequent but semantically important emotions.

To mitigate this issue, data augmentation techniques are commonly employed to increase the diversity and quantity of samples associated with underrepresented emotions [8], [15]. Augmentation strategies range from simple lexical substitutions to more advanced contextual generation methods that preserve semantic consistency and emotional intent [8], [15]. By enriching minority classes and exposing models to a broader range of emotional expressions, data augmentation aims to improve generalization performance and robustness in imbalanced multi-label classification scenarios [9].

III. METHODS

A. Research Design and Workflow

This study adopts a comparative experimental approach to evaluate the efficacy of lightweight Transformer-based models against a Convolutional Neural Network (CNN) baseline for multi-label emotion classification. The research workflow consists of four primary stages: (1) Data Preparation (2) Model Development, and (3) Performance Evaluation.

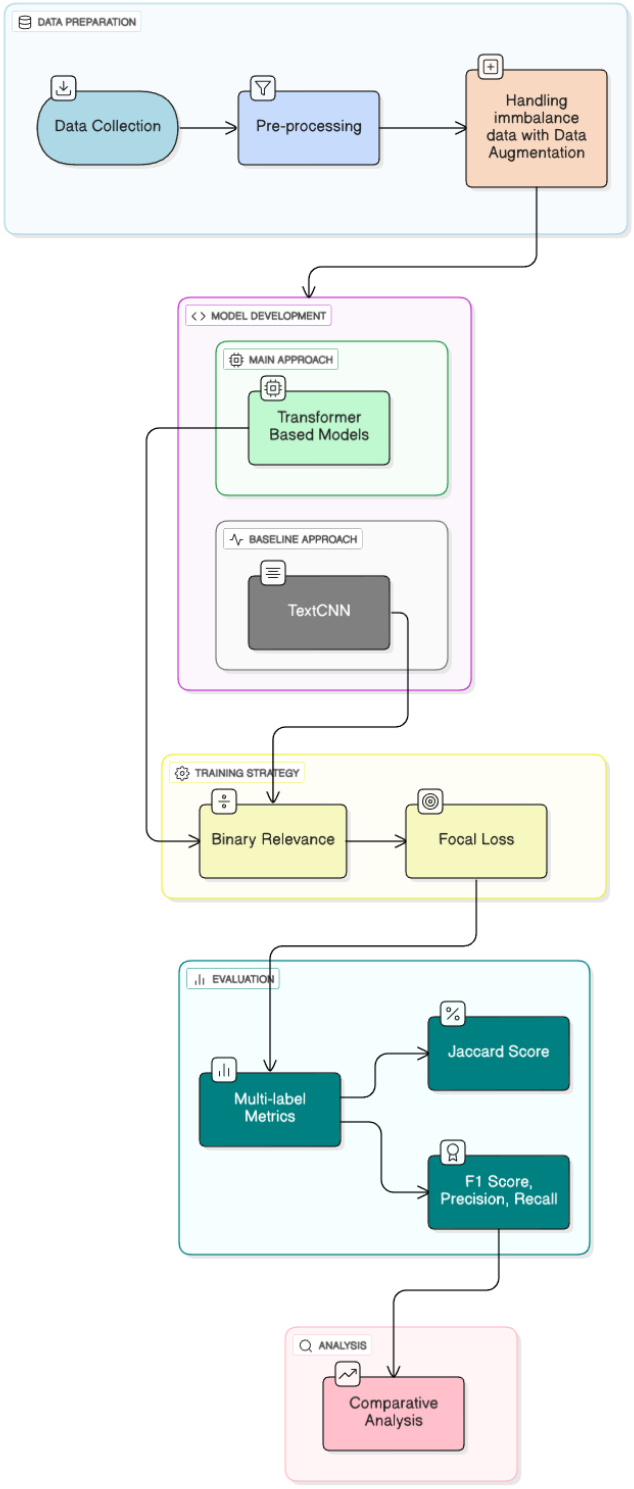

Fig. 1. Research Flow Diagram

B. Dataset and Data Pre-processing

Dataset utilized is the Shared Task of Emotion Recognition (INFO 557 Fall 2025), which classifies text into seven fine-grained emotion categories: admiration, amusement, gratitude, love, pride, relief, and remorse. Given the noisy nature of social media text, we employed a rigorous pre-processing pipeline inspired by established methodologies [16, 17].

C. Noise Removal and Normalization

Raw text often contains non-semantic elements that introduce noise into the learning process. Following the preprocessing standards set by [17, 18], we utilized regular expressions and specific libraries (such as ekphrasis) to clean the data. Steps included: Token Replacement: User mentions and URLs were replaced with generic tokens (e.g., <USER>, <URL>) to reduce vocabulary sparsity while preserving structural context [18]. Hashtag and Case Normalization: Hashtags were segmented to retain semantic meaning (e.g., converting #happy to happy), and all text was converted to lowercase to standardize the input [19]. Sub-word Tokenization: For Transformer-based models (DistilBERT, DistilRoBERTa, ALBERT), text was tokenized using model-specific sub-word tokenizers. This approach effectively handles Out-Of-Vocabulary (OOV) words by breaking them down into meaningful sub-word units, ensuring robust embedding generation [16, 18].

D. Handling Class Imbalance

Preliminary analysis of the dataset revealed a long-tail distribution, where minority emotion classes such as “Pride” and “Relief” were significantly underrepresented. Such imbalance often biases models toward majority classes. To mitigate this, we implemented Data Augmentation strategies. Aligned with the methods proposed by [18], we utilized oversampling techniques during the training phase to balance the ratio of positive to negative samples. This strategy aims to enrich the semantic diversity of minority classes without altering the fundamental emotional intent, thereby improving the model's recall for underrepresented labels.

E. Model Architectures

We compared a classical deep learning baseline against state-of-the-art efficiency-optimized Transformers.

1) TextCNN (Convolutional Neural Network)

TextCNN models have demonstrated strong performance in multi-label emotion classification across various domains, including social media, software engineering, and group sentiment analysis. For example, using TextCNN with word embeddings and hyperparameter optimization achieved F1-Micro scores of 84.6% on Jira and 76.9% on Stack Overflow datasets, outperforming previous methods in software engineering texts [20].

TextCNN as a strong baseline due to its proven capability in extracting local semantic features (n-grams) through convolutional filters of varying kernel sizes. Despite its simplicity compared to recurrent networks, CNNs have demonstrated high efficiency in capturing key emotional triggers within short texts.

2) Efficient Transformers (DistilBERT, DistilRoBERTa, ALBERT)

Transfer Learning approach by fine-tuning pre-trained Transformer models that have been optimized for efficiency: DistilBERT & DistilRoBERTa: Both are distilled versions of BERT and RoBERTa, respectively, using knowledge distillation to reduce parameters and speed up inference. DistilBERT, for example, is about 40% smaller and 60% faster than BERT, while maintaining 95–97% of its accuracy on tasks like GLUE and sentiment analysis [21]. ALBERT achieves efficiency by sharing parameters across layers and factorizing embeddings, resulting in a much smaller memory footprint. However, aggressive quantization or further compression can sometimes lead to notable performance drops, especially on already compact models [22]. These models were selected for their ability to model global context via self-attention mechanisms, which is critical for identifying co-occurring emotions in multi-label settings.

E. Training Strategy

The multi-label classification task was transformed into a set of binary classification problems, where the probability of each emotion label is predicted independently using a Sigmoid activation function at the output layer. Loss function is used to further address class imbalance, we explored the use of Focal Loss in addition to standard Binary Cross-Entropy.

$$L_{fl} = \left\{\frac{-(1-y)^{\gamma}\log(y), if\ y=1}{-y^{\gamma}\log(1-y), if\ y=0}\right\} \quad (1)$$

As demonstrated by Lin et al., the focal loss in (1) down-weights easy negatives and focuses the model's training on hard-to-classify examples, which is crucial for improving performance on minority emotions. The models were optimized using the AdamW optimizer with a decaying learning rate to ensure stable convergence during fine-tuning.

F. Evaluation Metrics

Standard accuracy is often insufficient for imbalanced multi-label datasets. Therefore, we employed the following metrics, consistent with SemEval competitions and recent literature [16, 19].

1) Jaccard Index (Multi-label Accuracy)

The Jaccard index measures the intersection over the union of the predicted and ground-truth label sets, providing a strict measure of exact overlap [16], as defined in (2).

$$Jaccard = \frac{1}{|T|}\sum_{t\in T}\left|\frac{G_t \cap P_t}{G_t \cup P_t}\right| \quad (2)$$

2) F1 Score

Calculates the F1 score globally by counting total True Positives, False Negatives, and False Positives. This metric is useful for assessing overall system performance.

3) Precision

Precision measures the ratio of correctly predicted positive labels to the total predicted positive labels, as defined in (3). In a multi-label context, high precision indicates a low rate of false positives (i.e., the model does not over-predict incorrect labels).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

4) Recall

Recall measures the ratio of correctly predicted positive labels to the total actual positive labels in the ground truth, as defined in (4). High recall indicates the model's sensitivity in capturing relevant emotion labels, which is critical for minority classes.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

IV. Results and Discussions

Based on the objectives and methodologies outlined in the previous sections, this section presents the experimental results and comparative analysis of the proposed models for multi-label emotion classification. We evaluate the performance of TextCNN and several efficient Transformer-based architectures—namely DistilBERT, DistilRoBERTa, and ALBERT—under a consistent multi-label learning framework. The results are reported using standard multi-label evaluation metrics to assess both overall classification effectiveness and per-label performance, with particular attention to minority emotion classes affected by data imbalance. Through this analysis, we aim to highlight the strengths and limitations of each model and to examine the impact of the adopted imbalance-handling strategies on improving classification robustness.

A. DistilBERT

Table 1 shows a detailed comparison of the DistilBERT performance before and after data augmentation for each emotion label. The overall Micro-F1 score increases from 0.8497 to 0.8525, indicating a slight improvement in aggregate multi-label performance. For Admiration, the F1-score changes marginally from 0.810 to 0.807, with a small decrease in precision and a slight increase in recall. Amusement shows a nearly identical performance across both settings, with the F1-score increasing slightly from 0.866 to 0.867. Gratitude exhibits a clear improvement, where the F1-score rises from 0.926 to 0.931 due to higher recall despite a small drop in precision.

TABLE I. Performance Comparison of DistilBERT Before and After Data Augmentation

Emotion Label	Before Augmentation			After Augmentation		
	P	R	F1	P	R	F1
Admiration	0.830	0.791	0.810	0.812	0.803	0.807
Amusement	0.858	0.875	0.866	0.851	0.884	0.867
Gratitude	0.947	0.905	0.926	0.923	0.939	0.931
Love	0.782	0.897	0.835	0.791	0.917	0.849
Pride	0.727	0.533	0.615	0.875	0.467	0.609
Relief	0.429	0.333	0.375	0.300	0.167	0.214
Remorse	0.870	0.882	0.876	0.859	0.897	0.878

P = Precision; R=Recall; F1=F1-score

For Love, the F1-score increases from 0.835 to 0.849, accompanied by improvements in both precision and recall. In the case of Pride, precision increases substantially from 0.727 to 0.875, while recall decreases from 0.533 to 0.467, resulting in a relatively stable F1-score (0.615 to 0.609). Relief shows lower metric values after augmentation, with the F1-score decreasing from 0.375 to 0.214 as both precision and recall decline. Finally, Remorse demonstrates a slight improvement, with the F1-score increasing from 0.876 to 0.878, driven by a higher recall value.

Figure 2 illustrates the confusion matrix of the DistilBERT model evaluated on the augmented dataset, which yields the best overall performance compared to the non-augmented

setting. The most notable strengths are observed for Gratitude and Remorse, where the model achieves high true positive counts (505 and 701, respectively) while maintaining very low false positives (6 for Gratitude and 3 for Remorse). This indicates that, after augmentation, the model is able to capture discriminative patterns for these emotions with high reliability and minimal confusion with other classes.

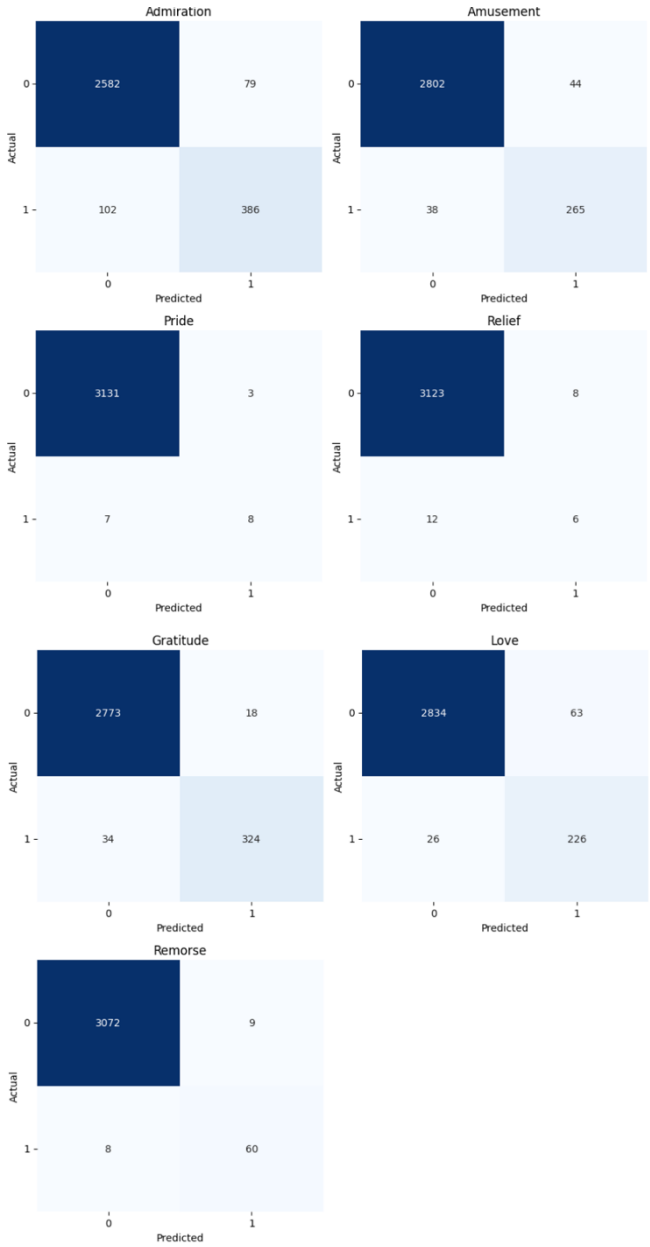

Fig. 2. Confusion Matrix of DistilBERT Model (Best Model: Using Data Augmentation)

In contrast, the confusion matrix highlights remaining challenges for certain minority emotions, particularly Relief and Pride. Although data augmentation enables the model to correctly identify a substantial number of Relief instancesΩ(336 true positives), this improvement is accompanied by a large number of false positives (588), suggesting a tendency toward overprediction. A similar trade-off is observed for Pride, with 124 true positives and 146 false positives. These findings indicate that, while data augmentation significantly enhances the model's sensitivity to underrepresented emotions, further refinement is needed to reduce misclassification and improve precision for these labels.

B. DistilRoBERTa

Table 2 shows the performance of the DistilRoBERTa model before and after data augmentation for each emotion label. Overall, the Micro-F1 score decreases from 0.8426 before augmentation to 0.8224 after augmentation. At the label level, data augmentation leads to improved F1-scores for several emotions, including Admiration, Amusement, Gratitude, Love, and Remorse, indicating better precision–recall balance for these classes. Notably, Amusement and Love show substantial gains after augmentation.

In contrast, the effect of augmentation varies for minority classes. Pride experiences a decrease in F1-score despite an increase in recall, while Relief shows a marked improvement in recall and F1-score after augmentation. These results illustrate that, although data augmentation enhances performance for certain emotion categories, its overall impact on DistilRoBERTa is mixed, as reflected by the lower Micro-F1 score after augmentation.

TABLE II. PERFORMANCE COMPARISON OF DISTILROBERTA BEFORE AND AFTER DATA AUGMENTATION

Emotion Label	Before Augmentation			After Augmentation		
	P	R	F1	P	R	F1
Admiration	0.778	0.820	0.798	0.852	0.859	0.856
Amusement	0.844	0.875	0.859	0.965	0.935	0.950
Gratitude	0.953	0.897	0.924	0.945	0.947	0.946
Love	0.772	0.901	0.832	0.903	0.903	0.903
Pride	0.727	0.533	0.615	0.465	0.703	0.560
Relief	0.400	0.222	0.286	0.398	0.876	0.547
Remorse	0.881	0.868	0.874	0.974	0.858	0.913

P = Precision; R=Recall; F1=F1-score

Figure 3 illustrates the confusion matrices of the DistilRoBERTa model before data augmentation. The results show that the model achieves strong classification performance for the majority emotion categories, including Admiration, Amusement, Gratitude, Love, and Remorse, as evidenced by the high concentration of values along the main diagonal. This pattern indicates that most samples are correctly classified, with relatively low rates of false positives and false negatives. Such behavior reflects a well-calibrated model that effectively captures dominant emotional patterns in the original dataset, contributing to the relatively high overall performance observed prior to augmentation.

After data augmentation, the performance of DistilRoBERTa declines. This reduction can be attributed to changes in the data distribution introduced by the augmented samples, which may differ from the original linguistic and emotional characteristics of the dataset. The inclusion of augmented instances can affect the balance between precision and recall and disrupt previously optimized decision thresholds, leading to an increase in misclassifications across labels. As a result, although augmentation may improve coverage for certain minority emotions, the cumulative effect of additional prediction errors can lead to a lower Micro-F1 score at the global level.

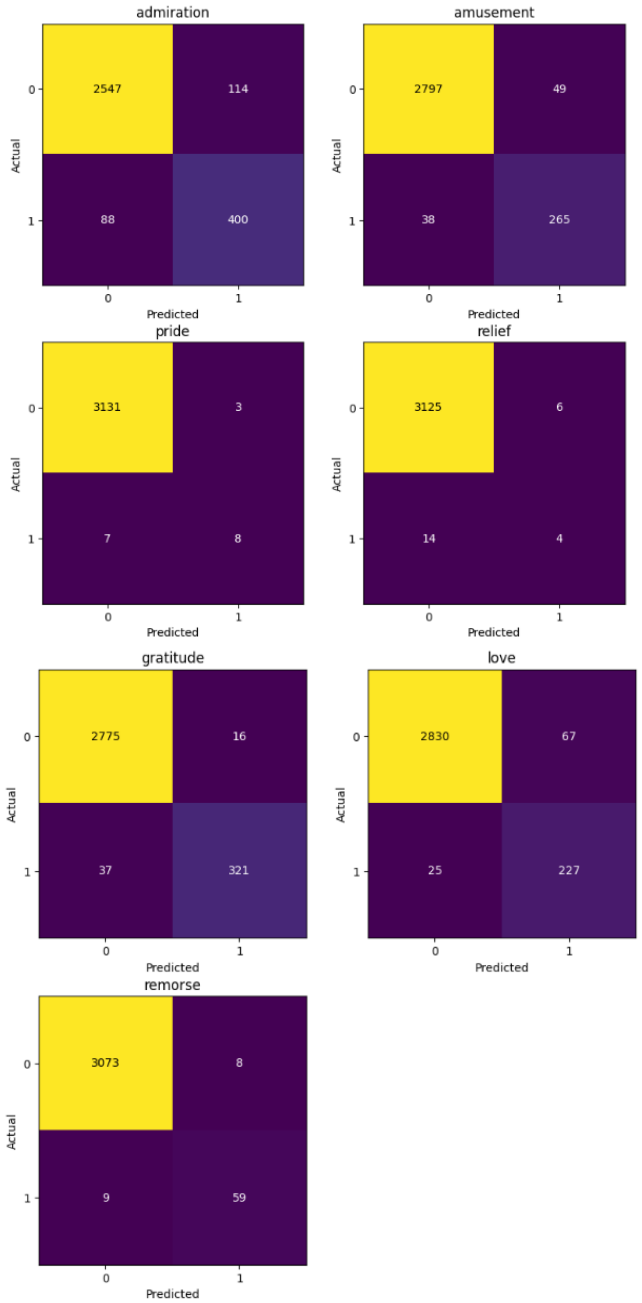

Fig. 3. Confusion Matrix of DistilRoBERTa Model (Best Model: Without Data Augmentation)

C. ALBERT

Table 3 presents the performance of the ALBERT model before and after data augmentation for emotion classification. Overall, data augmentation does not uniformly improve performance across all emotion labels. For more common emotions such as Admiration, Amusement, Gratitude, and Love, F1-scores decrease, suggesting that augmented data may introduce noisy or less representative examples. This indicates that simply increasing data volume does not necessarily enhance recognition of well-represented emotions and can sometimes reduce precision.

Conversely, augmentation shows clear benefits for rarer or highly variable emotions. Pride exhibits slightly higher recall but lower precision, while Relief and Remorse achieve substantial gains in recall, allowing the model to detect more instances of these emotions. However, precision remains moderate, resulting in only partial improvements in F1-scores. These results highlight a trade-off between precision and recall: augmentation can help capture underrepresented emotions, but careful selection and balancing of augmented data are essential to maintain overall model performance.

TABLE III. PERFORMANCE COMPARISON OF ALBERT BEFORE AND AFTER DATA AUGMENTATION

Emotion Label	Before Augmentation			After Augmentation		
	P	R	F1	P	R	F1
Admiration	0.770	0.812	0.791	0.717	0.540	0.616
Amusement	0.823	0.900	0.860	0.818	0.574	0.675
Gratitude	0.940	0.880	0.909	0.943	0.651	0.770
Love	0.825	0.895	0.858	0.791	0.534	0.638
Pride	0.625	0.556	0.588	0.472	0.591	0.525
Relief	0.341	0.452	0.389	0.314	0.749	0.442
Remorse	0.822	0.855	0.838	0.387	0.880	0.538

P = Precision; R=Recall; F1=F1-score

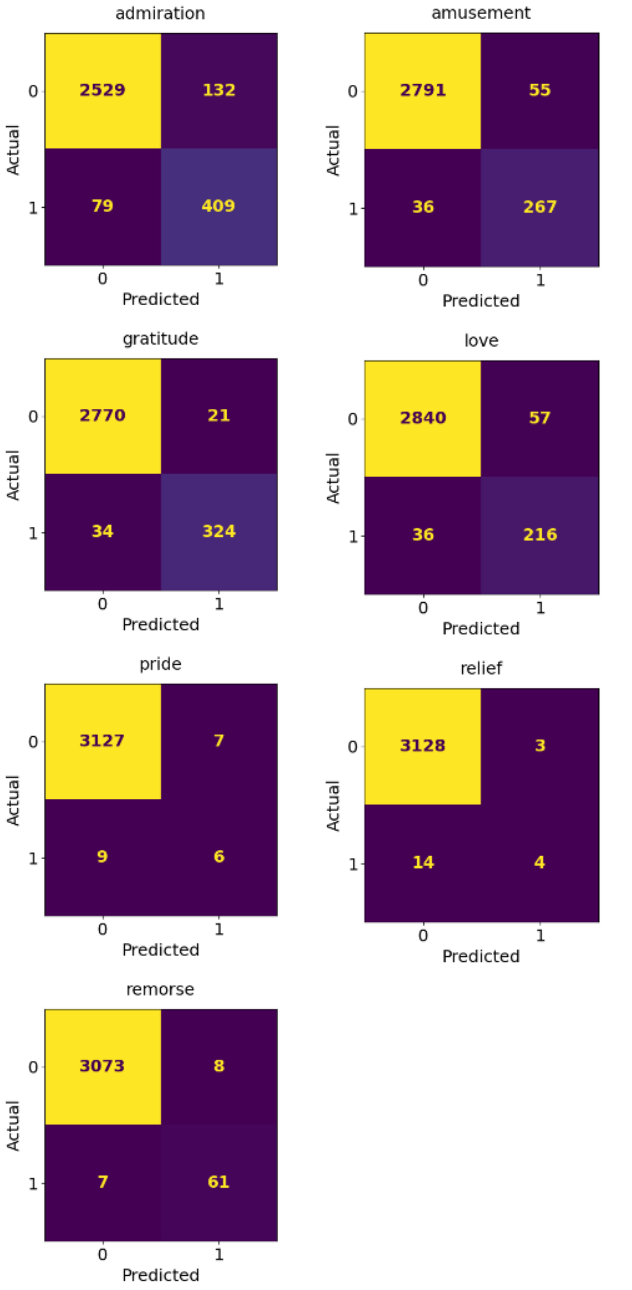

Fig. 4. Confusion Matrix of ALBERT Model (Best Model: With Data Augmentation)

Figure 4 presents the confusion matrix of the ALBERT model using data augmentation. The matrix shows that the model accurately predicts the more frequent emotions such as Admiration, Amusement, Gratitude, and Love, with high true positive counts and relatively few misclassifications. This indicates that data augmentation effectively enhances the model's ability to recognize these emotions.

For less frequent emotions like Pride and Relief, true positive counts are much lower, and some misclassifications remain, suggesting that detecting rare emotions is still challenging despite augmentation. Overall, the confusion matrix illustrates that data augmentation improves overall classification performance, particularly for common emotions, while highlighting the ongoing difficulty in accurately predicting less frequent emotional labels.

D. TextCNN

Based on the results shown in the table, the impact of data augmentation on the TextCNN model varies across emotion labels. For admiration, the F1-score increases from 0.717 to 0.787, indicating improved balance between precision and recall after augmentation. A similar positive trend is observed for amusement, gratitude, love, and remorse, all of which achieve higher F1-scores after augmentation. In particular, gratitude and remorse show substantial gains, reflecting the model's improved ability to correctly identify these emotions when exposed to more diverse training examples.

In contrast, the performance for pride and relief remains relatively challenging. Although recall improves notably for these two labels after augmentation, this improvement is accompanied by a considerable drop in precision, resulting in lower F1-scores compared to other emotions. At the overall level, these trade-offs contribute to a slight decrease in Micro F1 from 0.796 before augmentation to 0.789 after augmentation, suggesting that the global performance of TextCNN does not consistently benefit from data augmentation. These results indicate that while augmentation can enhance recognition for certain emotion categories, its effectiveness is label-dependent and may introduce noise for emotions that are inherently ambiguous or underrepresented.

TABLE IV. PERFORMANCE COMPARISON OF TEXTCNN BEFORE AND AFTER DATA AUGMENTATION

Emotion Label	Before Augmentation			After Augmentation		
	P	R	F1	P	R	F1
Admiration	0.700	0.736	0.717	0.764	0.811	0.787
Amusement	0.824	0.772	0.797	0.972	0.838	0.900
Gratitude	0.937	0.908	0.922	0.988	0.900	0.942
Love	0.804	0.813	0.809	0.908	0.873	0.890
Pride	0.750	0.200	0.316	0.459	0.593	0.518
Relief	0.667	0.111	0.190	0.364	0.745	0.489
Remorse	0.871	0.794	0.831	0.996	0.842	0.912

P = Precision; R=Recall; F1=F1-score

Figure 5 presents the confusion matrix of the TextCNN model after data augmentation and illustrates a clear change in the model's prediction behavior, particularly for minority emotion classes. Compared to the pre-augmentation setting, the augmented model produces a substantially higher number of true positive predictions for Relief and Pride, indicating that these emotions, which were previously rarely detected, are now more consistently recognized. This suggests that data augmentation effectively mitigates the model's tendency to favor the majority negative class and improves its sensitivity to underrepresented emotions.

However, Figure 5 also reveals an increase in false positive predictions for these minority categories. The improved detection of Relief and Pride is accompanied by a higher number of samples from other emotions being incorrectly assigned to these labels, reflecting a trade-off between sensitivity and specificity. Overall, the confusion matrix demonstrates that data augmentation shifts the TextCNN model toward a more balanced classification behavior, while also highlighting the remaining challenge of reducing false positive errors for minority emotion labels.

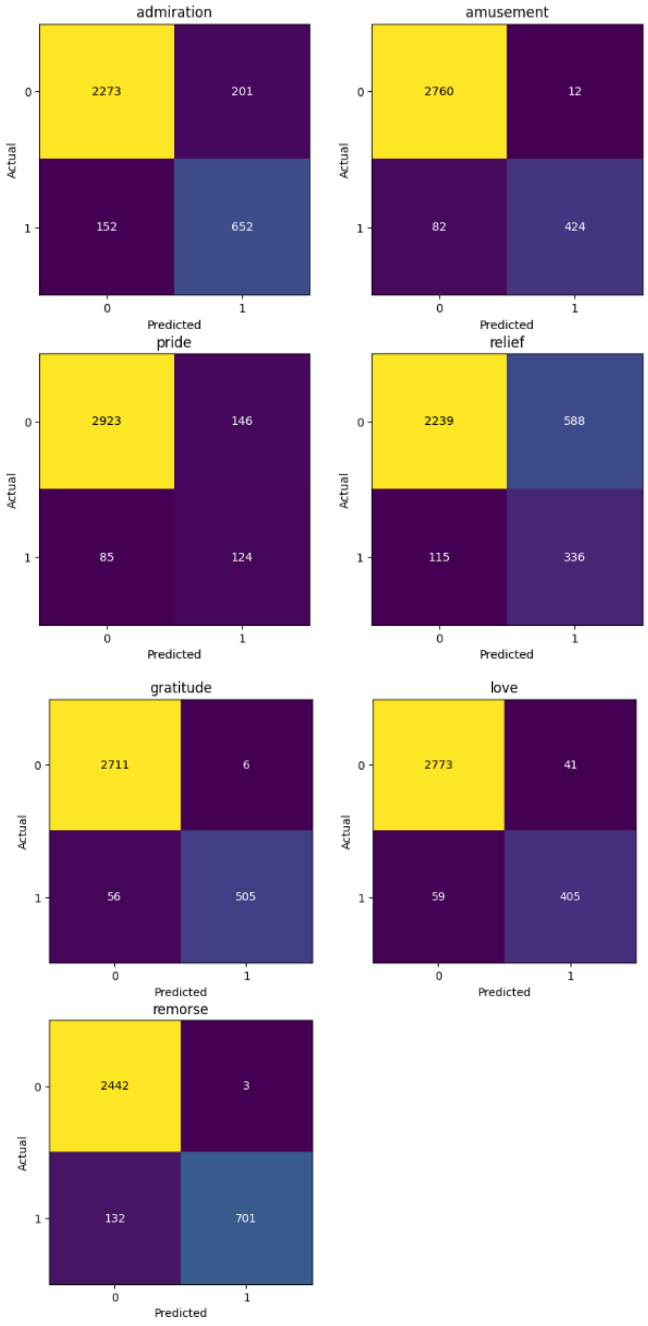

Fig. 5. Confusion Matrix of TextCNN Model (Best Model: Using Data Augmentation)

Based on the comparative evaluation of four models—DistilBERT, DistilRoBERTa, ALBERT, and TextCNN—DistilBERT demonstrates the best overall performance for the multi-label emotion classification task. As shown in Table I, it achieves the highest Micro-F1 score and maintains relatively balanced F1-scores across most emotion

categories, highlighting its superior ability to capture contextual and emotional information in text. The confusion matrix on the augmented dataset (Fig. 1) further confirms this, with high true positive counts and minimal false positives for key emotions such as Gratitude and Remorse.

Comparing performance before and after data augmentation, DistilBERT shows slight overall improvements with some label-specific differences. F1-scores for Gratitude, Love, and Remorse increase modestly due to higher recall, indicating improved sensitivity to less frequent or more subtle emotional expressions. For example, Gratitude rises from 0.926 to 0.931, Love from 0.835 to 0.849, and Remorse from 0.876 to 0.878. Common emotions such as Admiration and Amusement remain largely stable, with F1-scores of 0.810 to 0.807 and 0.866 to 0.867, respectively, reflecting robustness to changes in the training data distribution.

Minority emotions such as Pride and Relief show mixed effects. Pride exhibits higher precision but slightly lower recall, resulting in a relatively stable F1-score (0.615 to 0.609), whereas Relief gains substantial recall at the cost of precision, causing its F1-score to decrease from 0.375 to 0.214. These patterns, also visible in the confusion matrices, highlight the trade-off between sensitivity and specificity: augmented data enhances detection of underrepresented emotions but may introduce more false positives.

Overall, the before-and-after comparison demonstrates that data augmentation is particularly beneficial for improving recall on rare or highly variable emotions, while its effect on well-represented emotions is limited. DistilBERT, leveraging both its efficient Transformer architecture and targeted augmentation, achieves the most reliable and balanced performance among the evaluated models. The label-specific F1 trends and confusion matrix analysis collectively emphasize the importance of carefully designing augmentation strategies to optimize detection while maintaining precision across all emotion categories.

IV. Conclusion

This study evaluated TextCNN and three efficient Transformer-based models—DistilBERT, DistilRoBERTa, and ALBERT—for multi-label emotion classification under imbalanced data conditions. Overall, Transformer-based models outperform the CNN baseline, highlighting the importance of contextualized representations for capturing subtle and overlapping emotional expressions. Data augmentation improves recall and sensitivity for several minority emotions, such as Gratitude, Love, and Remorse, but may also introduce false positives, leading to mixed effects on overall performance for some models.

Among all models, DistilBERT achieves the best overall performance, with the highest Micro-F1 and relatively stable per-label results across both common and rare emotions. Augmentation further strengthens its detection of underrepresented classes, while maintaining robustness for frequent labels. Although minority emotions like Pride and Relief show mixed outcomes, DistilBERT benefits most consistently from augmentation compared to DistilRoBERTa, ALBERT, and TextCNN. These findings indicate that combining an efficient Transformer architecture with carefully designed data augmentation provides the most reliable and effective approach for multi-label emotion classification in imbalanced settings.

Appendix

The dataset, code, and research summary (poster format) used in this study for emotion classification experiments are publicly available on Github: https://github.com/herlinalim-ugm/Final-Project-PMML.git. Researchers can access the repository to reproduce the results or adapt the implementation for further studies.

References

[1] N. Lin, S. Fu, X. Lin, and L. Wang, "Multi-label emotion classification based on adversarial multi-task learning," *Inf Process Manag*, vol. 59, no. 6, Nov. 2022, doi: 10.1016/j.ipm.2022.103097.

[2] J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel, "Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018, pp. 1097–1102.

[3] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent Emotion Memory for Multi-Label Emotion Classification," in *Proceedings of the AAAI conference on artificial intelligence*, 2020, pp. 7692–7699. [Online]. Available: www.aaai.org

[4] C. Huang, A. Trabelsi, X. Qin, N. Farruque, L. Mou, and O. Zaïane, "Seq2Emo: A Sequence to Multi-Label Emotion Classification Model," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2021, pp. 4717–4724. [Online]. Available: https://github.com/

[5] I. Ameer *et al.*, "Multi-Label Emotion Classification in Texts Using Transfer Learning," *Expert Syst Appl*, vol. 213, 2023, [Online]. Available: https://en.oxforddictionaries.com/definition/emotion.

[6] M. Amirhosseini, N. Berardinelli, K. Gaikwad, C. Iwuchukwu, and M. Ahmed, "Integrated Sentiment and Emotion Analysis of the Ukraine-Russia Conflict Using Machine Learning and Transformer Models," in *Proceedings of the 14th International Conference on Data Science, Technology and Applications*, SCITEPRESS - Science and Technology Publications, 2025, pp. 191–202. doi: 10.5220/0013645500003967.

[7] X. Guo and B. Eng, "MULTI-LABEL CLASSIFICATION AND SENTIMENT ANALYSIS ON TEXTUAL RECORDS," MCMASTER UNIVERSITY, 2019.

[8] X. Wei, J. Huang, R. Zhao, H. Yu, and Z. Xu, "Multi-Label Text Classification Model Based on Multi-Level Constraint Augmentation and Label Association Attention," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, Jan. 2024, doi: 10.1145/3586008.

[9] Y. Chai *et al.*, "Compositional Generalization for Multi-Label Text Classification: A Data-Augmentation Approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 17727–17735. [Online]. Available: https://github.com/yychai74/LD-VAE.

[10] C. Huang, A. Trabelsi, X. Qin, N. Farruque, and O. R. Za¨ıaneza¨ıane, "Seq2Emo for Multi-label Emotion Classification Based on Latent Variable Chains Transformation," *arXiv preprint arXiv:1911.02147*, 2019, [Online]. Available: www.aaai.org

[11] H. He and R. Xia, "Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2018, pp. 250–259.

[12] H. Mulki, C. B. Ali, H. Haddad, and I. Babaoˇ Glu, "Tw-StAR at SemEval-2018 Task 1: Preprocessing Impact on Multi-label Emotion Classification," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 167–171. [Online]. Available: http://snowball.tartarus.org/algorithms/english/stop.txt

[13] X. Tian, Y. Qin, R. Huang, and Y. Chen, "A Label Information Aware Model for Multi-label Text Classification," *Neural Process Lett*, vol. 56, no. 5, Oct. 2024, doi: 10.1007/s11063-024-11692-z.

[14] M. Hasan, E. Rundensteiner, and E. Agu, "DeepEmotex: Classifying Emotion in Text Messages using Deep Transfer Learning," in *IEEE international conference on big data*, IEEE, 2022, pp. 5143–5152. Accessed: Dec. 11, 2025. [Online]. Available: https://arxiv.org/abs/2206.06775

[15] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, "Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages," *Sustainability (Switzerland)*, vol. 15, no. 16, Aug. 2023, doi: 10.3390/su151612539.

[16] N. Lin, S. Fu, X. Lin, and L. Wang, "Multi-label emotion classification based on adversarial multi-task learning," *Information Processing and Management*, vol. 59, no. 6, p. 103097, 2022.

[17] M. Jabreel and A. Moreno, "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets," *Applied Sciences*, vol. 9, no. 6, p. 1123, 2019.

[18] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Systems With Applications*, vol. 213, p. 118534, 2023.

[19] N. Ashraf, L. Khan, S. Butt, H. Chang, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification of Urdu tweets," *PeerJ Computer Science*, vol. 8, p. e896, 2022.

[20] A. A. Wagan and S. Li, "Multilabeled Emotions Classification in Software Engineering Text Using Convolutional Neural Networks and Word Embeddings," *Journal of Software: Evolution and Process*, vol. 37, 2025, doi: 10.1002/smr.70010

[21] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowledge and Information Systems*, vol. 66, 2024, doi: 10.1007/s10115-024-02214-3.

[22] K. Kim and C. Jeong, "F-ALBERT: A Distilled Model from a Two-Time Distillation System for Reduced Computational Complexity in ALBERT Model," *Applied Sciences*, vol. 13, no. 17, p. 9530, 2023, doi: 10.3390/app13179530.