CONFOBCE-MLTC: A NOVEL CONTRASTIVE FOCAL BINARY CROSS ENTROPY LOSS FUNCTION FOR MULTI-LABEL TEXT-BASED EMOTION CLASSIFICATION

Hassan Adamu [1], Masrah Azrifah Azmi Murad [2*], Nurul Amelina Nasharuddin[3]

[1] Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

[2] Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

[3] Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Email: gs66245@student.upm.edu.my[1], masrah@upm.edu.my[2*] (Corresponding author), nurulamelina@upm.edu.my[3]

ABSTRACT

Multi-label text classification remains challenging due to class imbalance, overlapping label distributions, and inter-label dependencies. Existing loss functions often address these issues independently, resulting in limited robustness and inconsistent generalization. To address these limitations, this paper proposes ConFoBCE-MLTC, a correlation-aware composite loss function that integrates binary cross-entropy, focal modulation, contrastive learning, and label correlation alignment within a unified optimization framework. The proposed method improves minority-class sensitivity, enhances feature separability, and explicitly models label co-occurrence structures without requiring architectural modifications. The approach was evaluated using multiple transformer-based architectures, including BERT, RoBERTa, DistilBERT, and mBERT, across both high-resource and low-resource datasets, including the newly constructed Hausa Emotion Corpus (HaEmoC_V1). Experimental results demonstrate consistent improvements over conventional loss functions and recent asymmetric loss formulations. In particular, mBERT combined with ConFoBCE-MLTC achieved the strongest overall performance, reaching 82.24 Micro-F1, 80.18 Macro-F1, and 73.31 Jaccard Score on the newly constructed dataset (HaEmoC_V1). Additional ablation and sensitivity analyses confirm the complementary contribution of focal weighting, contrastive representation learning, and label correlation alignment. These findings demonstrate the effectiveness of the proposed framework for multilingual multi-label emotion classification, particularly in low-resource settings.

Keywords: Multi-label emotion; Low-resource language; Transformer model; Contrastive loss; Focal loss; BCE loss; Hausa language.

1.0 INTRODUCTION

Emotion detection from text is a crucial task in natural language processing (NLP), playing a significant role in applications such as social media monitoring, disaster management, mental health assessment and customer sentiment analysis [1]. Understanding human emotions in textual data enables automated systems to analyse opinions, and facilitate effective decision-making in real-world situations [2]. Traditional sentiment analysis focuses on categorizing text into positive, negative, or neutral sentiments [3-4]. However, emotion classification goes beyond polarity, identifying emotional states such as joy, anger, sadness, fear, and trust [5-6]. In many real-world cases, a single text can express multiple emotions simultaneously, necessitating multi-label emotion classification. Recognizing clear emotions is essential for creating human-computer interaction systems that respond with empathy. In addition, accurate emotion detection supports digital well-being, enables more tailored recommendations, and helps identify potential conflicts in online communities.

Despite progress in text-based sentiment analysis and emotion classification, multi-label emotion detection presents several challenges, particularly for low-resource languages [7]. Many existing deep learning and machine learning models rely on large annotated datasets and rich linguistic resources, which are scarce for low-resource languages. Moreover, emotion labels in textual data are often imbalanced, with certain emotions appearing more

frequently than others, leading to biased model predictions [8]. Another major challenge lies in the diversity of cultural and linguistic expressions of emotion, which complicates annotation consistency. Addressing these challenges requires robust techniques that mitigate data scarcity and class imbalance, ensuring more reliable and inclusive NLP application.

Multi-label emotion detection extends single-label classification by allowing a text sample to be assigned multiple emotions rather than being restricted to one category [9]. This characteristic makes multi-label classification more complex, as emotions can overlap, creating ambiguities in automatic annotation and model training. For example, a disaster-related tweet might express both fear and sadness, while a post about social activism might contain anger and hope. Standard classification models struggle with such overlapping emotions, highlighting the need for specialized learning strategies [10]. These overlapping features also make annotation more subjective, as different annotators may interpret the same text with different emotional meanings. Therefore, reliable multi-label classification needs both advanced models and strong annotation methods to reduce uncertainty.

Recent advancements in transformer-based models, such as BERT, RoBERTa, mBERT, and DistilBERT, have significantly improved multi-label text classification by leveraging contextual embeddings [11]. However, fine-tuning these models for multi-label emotion detection requires effective loss functions that can balance frequent and rare emotions, enhance inter-class separability, and improve generalization. Existing loss functions like Binary Cross Entropy (BCE) fail to adequately address class imbalance and label dependencies, necessitating new techniques for better performance [12-13]. Moreover, the performance of these models heavily depends on high-quality annotated datasets, which remain scarce for low-resource languages. Incorporating domain adaptation strategies and transfer learning approaches can further boost robustness and transferability across diverse linguistic and cultural contexts.

The availability of large-scale annotated datasets greatly affects the performance of multi-label classification models [14]. High-resource languages such as English, French, and Chinese benefit from extensive labelled corpora, allowing deep learning models to generalize effectively on different tasks [15-16]. However, the success of fine-tuning transformer models in high-resource languages does not always extend to low-resource contexts, mainly due to limited annotated data [8]. Low-resource languages, including Hausa, Igbo, and Yoruba, often lack sufficiently large corpora for training deep learning models, which makes it difficult to build robust NLP applications [17-18]. As a result, transfer learning and fine-tuning techniques have become practical solutions, enabling models pre-trained on high-resource languages to adapt to low-resource contexts [19-20].

Loss functions are at the core of deep learning optimization, influencing how models learn and generalize. Effective loss functions play a crucial role in improving model robustness by enhancing discriminative learning, allowing models to differentiate between closely related emotions [21]. The standard loss functions such as Binary Cross Entropy (BCE) is widely used in multi-label classification but suffers from imbalanced label distributions [22]. Focal loss, originally proposed for object detection, addresses class imbalance by assigning higher weights to difficult-to-classify instances, improving model attention to minority labels [23-24]. Contrastive loss, on the other hand, enhances feature space separability by pulling similar instances closer together while pushing dissimilar instances apart, making it valuable for emotion recognition tasks [25-26]. Integrating these loss functions or designing hybrid approaches offers opportunities to capture the strengths of each while overcoming their weaknesses. Therefore, the choice of loss function becomes a determining factor in the overall success of multi-label emotion detection systems.

This study addresses these challenges through the following key contributions:

1. A proposed correlation-aware loss function (ConFoBCE-MLTC) that integrates binary cross-entropy, focal loss, and contrastive learning into a unified optimization framework, enabling simultaneous handling of class imbalance, feature separability, and label dependency modeling.
2. A label correlation alignment mechanism, based on Kullback–Leibler divergence, which explicitly enforces consistency between predicted and empirical label co-occurrence structures.
3. The construction of a new multi-label Hausa emotion dataset (HaEmoC_V1), providing a valuable benchmark for low-resource NLP research.
4. A comprehensive experimental evaluation across multiple transformer architectures and multilingual datasets, demonstrating consistent improvements over standard loss functions and validating the generalizability of the proposed approach.

In addition to proposing a unified loss framework, this study further investigates the robustness and practical applicability of the model through extensive sensitivity analysis, computational efficiency evaluation, and comparison with recent state-of-the-art multi-label loss functions. These additions ensure that the proposed method is not only effective but also reproducible and scalable across diverse linguistic settings.

The paper is structured into five sections; Section 2 provides a comprehensive review of related work in emotion detection, loss functions and transfer learning, establishing the foundational context for our research. Section 3 elaborates on the proposed hybrid loss function methodology, including its architecture and key

innovation. Section 4 presents the experimental setup, results, and a detailed analysis of the findings. Finally, Section 5 concludes the paper by discussing the limitations of the current study and outlining potential directions for future research.

2.0 RELATED WORKS

Sentiment analysis and emotion classification are foundational tasks in NLP but differ in scope and specificity. Sentiment analysis traditionally categorizes text into broad polarity classes (positive, negative, neutral) [27-29], while emotion detection identifies fine-grained affective states such as Ekman's six basic emotions (anger, joy, sadness, fear, surprise, disgust) [30] as shown in Fig. 1. The ability to detect emotions, a key part of affective computing, has become increasingly popular in recent years because it can be used in a variety of ways in fields like healthcare, marketing, and human-computer interaction [31]. It involves the recognition and categorization of human emotions, typically utilizing data such as facial expressions, speech intonations, or physiological signals like heart rate and skin conductance [32].

Fig. 1: Six basic emotions [30]

Emotion models serve as a cornerstone in this process, providing a structured framework for comprehending and classifying emotions. Prominent among these models is Geneva Emotion Wheel, devised by Klaus Scherer as shown in Fig. 2, offers a comprehensive alternative to basic emotions by extending the spectrum of emotions and accounting for cognitive appraisals, subjective experiences, and expressive behaviour [33].

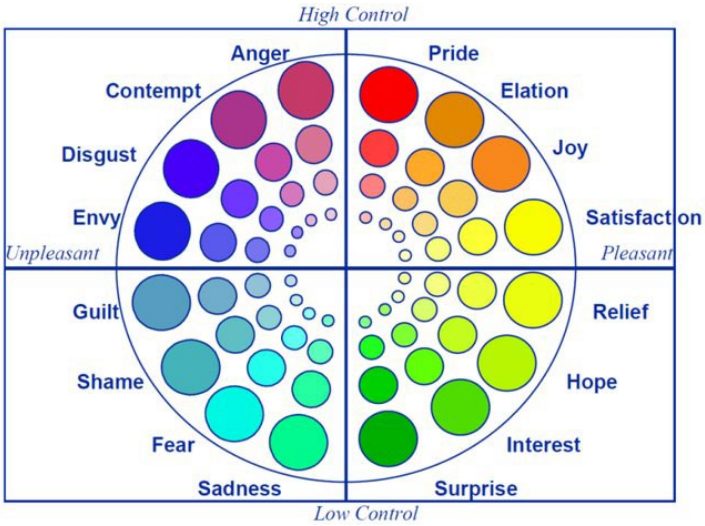

Fig. 2: Geneva emotion wheel [33].

The field of emotion detection is constantly evolving, due to the advances in deep learning, especially neural networks like CNNs and RNNs. These techniques have made emotion detection systems more accurate and adaptable. These models can efficiently process and analyse vast amounts of multimodal data, such as images, audio, and text, further enhancing their effectiveness in diverse real-world applications [34]. Table 1 presents different emotion models and theories.

Table 1: Emotion models and theories

Emotion Model	Type of Model	Number of States	Psychological State	Representations Discussion
Ekman's Six Basic Emotions [30]	Componential	6 (Happiness, Sadness, Anger, Fear, Disgust, Surprise)	Universal and biologically rooted basic emotions.	This model suggests that these basic emotions are universal across cultures, with distinct facial expressions and physiological responses.
Lazarus' Cognitive Appraisal Theory [35]	Componential	Multi-dimensional	Emotions result from cognitive appraisals of a situation.	Focuses on how individuals evaluate and interpret situations, leading to emotional responses that can vary based on these appraisals.
Plutchik's Wheel of Emotions [36]	Componential	8 Primary Emotions, Secondary Emotions	Primary and secondary emotions are two different types of emotions	Provides a visual representation of how different emotions are related to each other and can transform from one primary emotion to another.
Barrett's Theory of Constructed Emotion [37]	Constructivist	Multi-dimensional	Emotions are constructed from core affect, concepts, and situated conceptualizations.	This model emphasizes that emotions are not fixed categories but rather constructed based on an individual's past experiences and conceptualization of their emotional state.
Panksepp's Affective Neuroscience Theory [38]	Biological	Multi-dimensional	Emotions are associated with specific neural circuits in the brain.	This theory identifies the neural basis for primary emotional systems, such as Seeking, Fear, and Rage, and how they influence behaviour.

Multi-label frameworks have emerged to address overlapping emotions such as tweet expressing both joy and surprise, with datasets like GoEmotions [39-40] and SemEval-2018 Task 1 [41- 42], enabling models to predict multiple labels per text. However, label correlation and class imbalance remain persistent challenges, particularly in short or ambiguous texts [43-44]. For instance, [45] highlighted that rare emotions like "disgust" are often misclassified due to sparse training examples. Recent advances leverage transformer architectures like BERT and RoBERTa to capture contextual challenges in multi-label emotion detection. In a recent research by [46] demonstrated that BERT's self-attention mechanism improves label dependency modelling by dynamically weighting relevant tokens. While English datasets like EmoBank [47] provide rich annotations, languages like Hausa lack comparable resources, worsening model bias toward high-resource approach [48].

The role of linguistic and cultural context in emotion detection has also gained attention. For instance, Moussa *et al.* [49] emphasized that emotion lexicons developed for English often fail to capture culturally specific expressions in other languages, such as the Hausa term *fargaba* (a mix of anger and resentment). This misalignment complicates cross-lingual transfer learning. Recent work by [50] proposed culturally adaptive embeddings to mitigate this issue, but their focus on single-label classification limits utility in multi-label cases. Additionally, social media platforms introduce noise like sarcasm and emojis, which require specialized pre-processing. Chauhan *et al.*[51] introduced emoji-aware multitask framework to improve multimodal sarcasm detection, yet their framework remains untested in low-resource languages.

Low-resource languages face unique challenges, including code-switching (e.g., mixing Hausa and English in Nigerian social media) and dialectal variations [52-53]. Initiatives like MasakhaNEWS [54] aim to address these gaps by crowdsourcing annotations for African languages, but emotion detection remains underexplored. Recent

work on Yorùbá sentiment analysis [55] revealed that direct translation of English emotion lexicons fails to capture culturally specific terms. This highlights the need for linguistically and culturally tailored datasets like HaEmoC_V1, introduced in this work, to enable equitable NLP advancements.

Transfer learning leverages pre-trained models such as BERT trained on large datasets to extract general patterns [56], which are then fine-tuned on task-specific data like sentiment analysis [57]. This approach maximizes efficiency, particularly for low-resource languages or specialized tasks, by repurposing existing knowledge rather than training from scratch [58]. The general process of transfer learning modelling, as illustrated in Fig. 3, involves initializing with a pre-trained BERT model [59], which is trained on extensive datasets like an English corpus to capture universal language representations. This model is then fine-tuned [60], on task-specific data (e.g., sentiment analysis or emotion detection datasets) to adapt its parameters for specialized objectives. For instance, sentiment analysis models benefit from BERT's contextual embeddings, which are refined using labelled datasets to improve classification accuracy [61]. Similarly, emotion detection leverages the same framework, adjusting the model to recognize emotional signals [62]. This approach is particularly impactful for low-resource African languages, where labelled data is limited; fine-tuning pre-trained models enables effective adaptation without requiring large-scale datasets [52], By transferring knowledge from high-resource languages, the model bridges performance gaps, demonstrating the versatility of transfer learning in NLP [63].

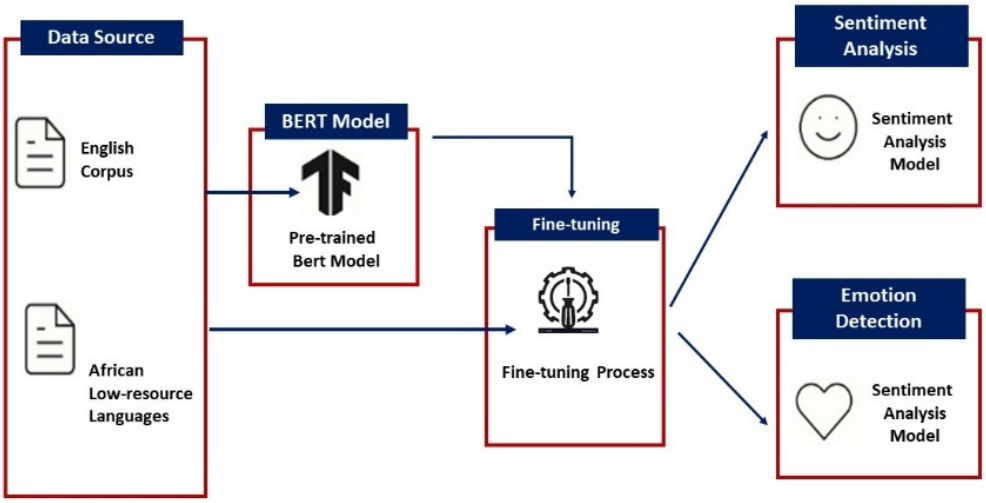

Fig. 3: General process of transfer learning

Transfer learning with pre-trained transformers has revolutionized NLP by enabling effective feature extraction from limited labelled data [64]. For multi-label classification, Binary Cross Entropy (BCE) is the default loss function but struggles with class imbalance and label dependencies [21]. In this study, Zheng et al. [65] addresses imbalance by down-weighting well-classified samples using Focal Loss, while Contrastive Loss [66] improves feature separation by clustering semantically similar embeddings. Hybrid approaches, such as adversarial learning with Contrastive Loss [67], have shown promise but lack mechanisms to handle both imbalance and label correlations simultaneously.

Recent innovations integrate domain-specific adaptations. For example, Li *et al.* [48] proposed a hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis classification but focused on single-label tasks. In social media emotion detection, Liang *et al.* [68] proposed an adaptive focal loss that dynamically adjusts class weights based on prediction difficulty. However, these methods neglect the relationship between label dependencies and cross-lingual transfer. Our ConFoBCE-MLTC unifies focal modulation for hard-sample mining and contrastive learning for label correlation modelling, offering a holistic solution for multi-label emotion detection across linguistic resource tiers.

The role of contrastive learning in multi-label contexts is still evolving. Traditional contrastive frameworks like SupCon [69] assume single-label data, forcing multi-label samples into disjoint clusters. To address this, Shining [70] proposed Label-Aware Contrastive Learning, which generates augmented views based on co-occurring labels. For example, a tweet labelled joy and surprise would be contrasted against samples sharing either label. Similarly, Huang *et al.* [71] introduced Partial Label Contrastive Learning for ambiguous samples but restricted experiments to synthetic data. Table 2 presents different studies on standard loss functions for both single and multi-label emotion classification across various NLP domains, such as computer vision, affective computing, and human-computer interaction.

Table 2: Comparative Analysis of Loss Functions in the contexts of Multi-Label Emotion Detection

Authors	Objectives	Methodology	Key Contributions	Weakness
[72]	Improve label dependency in multi-label emotion classification.	Contrastive learning with label-aware sampling.	Introduces contrastive learning for label correlations.	No BCE or focal loss integration.
[73]	Mitigate class imbalance in single-label text classification.	Joint optimization of focal + contrastive loss.	Combines focal and contrastive losses for imbalance.	Limited to single-label tasks.
[74]	Synthesize loss function advancements in NLP.	Survey of hybrid loss designs (focal, contrastive).	Reviews hybrid loss trends in NLP.	No BCE integration
[75]	Improve emotion recognition via multi-task learning.	Bidirectional GRU + attention + refined contrastive loss.	Multi-stage feature fusion; improved contrastive learning.	Computationally complex. No loss design.
[76]	Reduce negative sample impact in imbalanced multi-label datasets.	Asymmetric BCE loss to suppress easy negatives.	Down-weights negative gradients for imbalance.	No contrastive learning or focal loss integration.
[77]	Improve pseudo-labelling for semi-supervised multi-label learning.	Class-aware thresholds for pseudo-label assignment.	Class-distribution-aware thresholding strategy.	Focuses on margins, not loss design.
[78]	Enhance label separation in multi-label tasks.	Contrastive loss with hard negative mining.	Prioritizes challenging samples via hard negatives.	Ignores BCE and focal loss.
[79]	Benchmark transformer-based emotion detection.	Fine-tuning pre-trained models with BCE loss.	Establishes BCE-based Transformer baselines.	Uses vanilla BCE loss. No contrastive or focal loss integration
[80]	Capture label dependencies in multi-emotion detection.	GNNs modelling co-occurrence patterns.	Leverages graph networks for label correlations.	No loss function innovation.
[81]	Dynamically optimize loss weights for multi-label tasks.	Reinforcement learning (RL) for per-label weight adjustment.	RL-driven dynamic loss adaptation.	Computationally heavy. No any standard loss function.
[82]	Improve multi-label decision thresholds.	Adaptive threshold tuning via confidence scores.	Adjusts thresholds per label probabilistically.	Focuses on thresholds, not loss design.
[71]	Enhance sentiment extraction for POI recommendations.	BERT + contrastive learning + optimized loss.	Integrates sentiment into POI recommendations.	Not focused on multi-label emotion detection. No BCE or focal loss integration

Table 2: Continued.

Authors	Objectives	Methodology	Key Contributions	Weakness
[83]	Reduce modality dominance in multi-modal emotion recognition.	Hierarchical knowledge distillation (feature + label levels).	Mitigates shortcut learning via distillation.	Targets multi-modal loss, not emotion-specific. No loss design
[77]	Improve label semantics/correlations in MLTC.	Multi-perspective contrastive learning + label alignment.	Integrates label semantics via contrastive learning.	Not emotion-specific. No BCE or focal loss
[84]	Improve emotion detection in ambiguous social media text.	Attention + ensemble learning + noise handling.	Combines pre-processing and ensemble methods for imbalance/ambiguity.	Retains BCE loss, but ignore contrastive or focal loss functions
[85]	Design robust contrastive loss for multi-label tasks.	Gradient analysis + loss pruning.	Modifies contrastive loss via harmful gradient removal.	Focuses on pure contrastive loss optimization. No BCE or focal loss integration
[86]	Improve EEG-based emotion classification.	Contrastive learning for cross-subject EEG alignment.	Addresses EEG variability across domains.	Single-label focus. No focal or BCE loss integration
[87]	Adapt loss functions to few-shot multi-label tasks.	Meta-learning for loss parameter tuning.	Generalizable framework for few-shot settings.	Not emotion-specific. No loss design

The comparative analysis shown in Table 2 reveals distinct methodological approaches toward improving multi-label emotion classification, particularly in handling class imbalance and label dependencies. The proposed approach ConFoBCE-MLTC stands out by integrating focal loss, BCEWithLogit and contrastive learning into a unified loss function, whereas existing works often focus on one aspect. For instance, [72], [88] employ contrastive learning to enhance label separation but do not incorporate focal loss, limiting their ability to address class imbalance. Similarly, Xu *et al.* [89] combine focal and contrastive losses, but their study is constrained to single-label classification, making it less applicable to multi-label emotion detection. Other works, such as [90],[77], propose alternative loss function modifications (asymmetric BCE and class-aware thresholds) but lack contrastive learning components, reducing their effectiveness in capturing label relationships.

While some studies offer different contributions, they have limitations in terms of practical application and computational efficiency. Kang *et al.* [75] introduce a sophisticated multi-task learning approach combining bidirectional GRU, attention mechanisms, and contrastive loss refinement, yet the computational complexity may hinder real-world deployment. Likewise, Vairetti [81] applies reinforcement learning for dynamic loss weight optimization, but this approach is computationally expensive, making it less feasible for large-scale multi-label emotion datasets. On the other hand, works like [91-92]explore alternative techniques such as graph neural networks (GNNs) and multi-modal contrastive learning, which effectively model label dependencies but do not introduce innovations in loss function design.

Furthermore, certain studies focus on specific domains or tasks, while relevant, are not directly comparable to ConFoBCE-MLTC. For example, Hu *et al.* [86] apply contrastive learning to EEG-based emotion classification, which differs from text-based multi-label emotion detection. Similarly, Labib *et al.* [84] tackle ambiguity in social media emotion detection using ensemble learning and attention mechanisms but retain BCE loss without contrastive or focal loss integration. Chauhan et al. [51], introduce a meta-learning approach to adapt loss functions for few-shot tasks, making it broadly applicable but not tailored for emotion detection.

While several recent studies have explored hybrid or asymmetric loss formulations for multi-label classification, such as Asymmetric Loss (ASL), these approaches primarily focus on reweighting positive and negative samples without explicitly modelling label co-occurrence structures. In contrast, the proposed

ConFoBCE-MLTC integrates focal modulation, contrastive representation learning, and explicit label correlation alignment within a unified framework. Section 3 depicts the research methodology.

3.0 METHODOLOGY

Transformer-based models, introduced by [93], transformed natural language processing (NLP) through self-attention mechanisms and contextualized word representations. These architectures, illustrated in Fig. 4, enable models to capture long-range dependencies in text while maintaining computational efficiency, making them highly effective for diverse NLP tasks. Over the years, models such as BERT [59], RoBERTa [94], mBERT [59], and DistilBERT [95] have emerged as state-of-the-art solutions for multi-label text classification, including sentiment analysis and emotion detection. These transformer-based models are pre-trained on massive text corpora and later fine-tuned for domain-specific applications, allowing them to generalize effectively across various tasks [96]. Additionally, their ability to model contextual semantics enables better disambiguation of semantic-rich words, improving classification accuracy [97]. As a result, transformer-based models continue to drive advancements in NLP, offering unparalleled improvements in performance and adaptability [98].

In this work, we fine-tune multiple transformer-based models for multi-label emotion detection in both high-resource and low-resource languages. Given the challenges caused by limited labelled data in low-resource contexts, we employ innovative training strategies to enhance model robustness and generalization. One of the key challenges in multi-label classification is handling class imbalance and overlapping label distributions, which can significantly impact predictive performance. To address these issues, we propose the Contrastive Focal Binary Cross Entropy (ConFoBCE-MLTC) loss function, a hybrid approach that combines focal loss [92] and contrastive learning principles. Focal loss mitigates the effect of class imbalance by down-weighting well-classified instances and focusing on harder-to-classify examples [65], while contrastive learning enhances label discrimination by encouraging the model to learn more distinct feature representations [99].

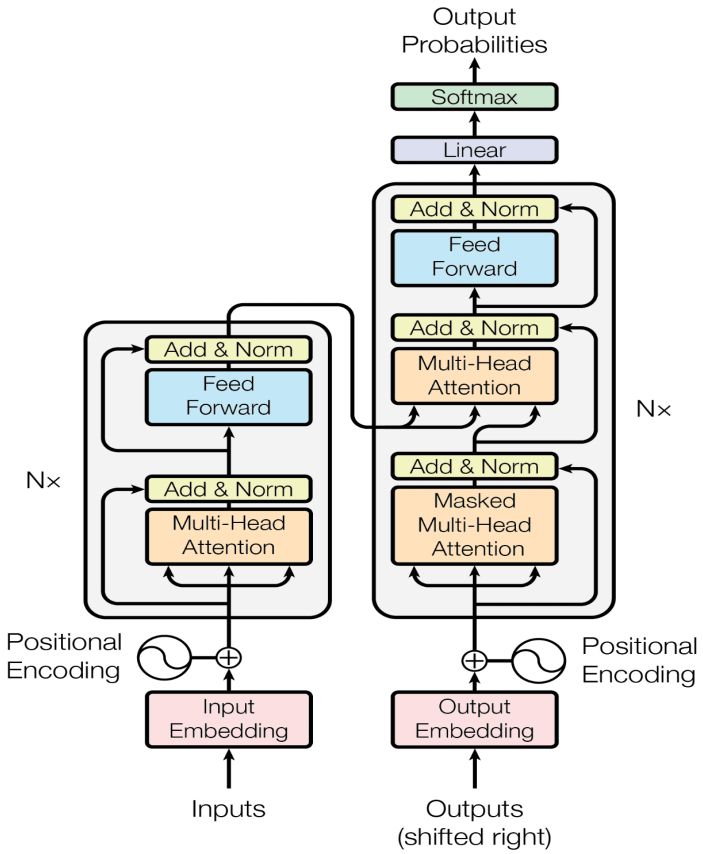

Fig. 4: Transformer architecture adapted from Vaswani et al. [93]

However, the current formulation does not explicitly model label dependencies, which are critical in multi-label emotion classification where emotions frequently co-occur. To address this limitation, we extend the proposed ConFoBCE-MLTC into a correlation-aware framework by incorporating (i) label-overlap-aware contrastive learning and (ii) a label co-occurrence alignment mechanism. This enables the model to capture both feature-level separability and label-level dependencies in a unified optimization objective. Section 3.1 described the selected transformer models architectures for this research.

3.1 Model Architectures

In this section, we describe the transformer-based models used for multi-label emotion detection for this work, highlighting their architecture, training strategies, and optimization techniques. Each model is fine-tuned with domain-specific datasets while incorporating the proposed technique, the hybrid ConFoBCE-MLTC loss function to enhance classification performance. We explore BERT, RoBERTa, mBERT, and DistilBERT, assessing their effectiveness in the contexts of both high-resource and low-resource languages.

A. *BERT:* We employ the 'bert-base-uncased' model on multi-label emotion datasets, fine-tuning it by adjusting batch size, learning rate, and the number of epochs. BERT is pre-trained using a masked language modelling (MLM) approach on the English Wikipedia and BookCorpus datasets. It consists of approximately 110M parameters [93].

B. *RoBERTa:* The RoBERTa model builds upon BERT by introducing dynamic masking, larger batch sizes, and extended training on a more extensive dataset. We implement the 'roberta-base' variant and fine-tune it on multi-label emotion detection corpora, setting the batch size and applying a learning rate. The RoBERTa model has demonstrated superior performance in various text classification tasks due to its enhanced training methodology [94].

C. *mBERT:* The 'bert-base-multilingual-cased' model, commonly referred to as mBERT, is trained on the top 104 languages with the largest Wikipedia corpora, including Hausa language. We fine-tune mBERT on both high and low-resource languages emotion datasets, leveraging the ConFoBCE-MLTC loss function to address imbalanced class distributions. The training process includes batch size tuning and dropout regularization for better generalization [59].

D. *DistilBERT:* We implement the 'distilbert-base-uncased' model, a distilled version of BERT, designed for efficiency with fewer parameters (approximately 66M) while maintaining high accuracy. The reduced model size enables faster inference, making it suitable for resource-constrained environments [95].

3.2 Dataset description

This research uses three datasets in both high-resource and low-resource languages to evaluate the effectiveness of the proposed hybrid loss function, the ConFoBCE-MLTC. These datasets cover diverse range of linguistic and cultural contexts, ensuring a comprehensive assessment of our approach. Table 3 provides an overview of the datasets, highlighting their language classification, dataset size, and sources. Among them, the Hausa Emotion Corpus (HaEmoC_V1) was specifically collected and constructed by us to address the scarcity of annotated Hausa emotion datasets for NLP tasks.

Table 3: Description of the datasets used in this work

Authors	Datasets	High-resource Language	Low-resource Language	Train	Development	Test	Total samples
[100]	SemEval-2018 Task 1 (Task E-c)	English	-	6838	886	3259	10,983
[101]	Indonesian Abusive and Hate Speech Twitter Text	-	Bahasa Indonesian	-	-	-	3,293
Current authors of this article	**Hausa Emotion Corpus (HaEmoC_V1)**	**-**	**Hausa**	**-**	**-**	**-**	**12,811**

3.3 Data acquisition and development of HaEmoC_V1

This section describes the dataset acquisition and organization process for the Hausa Emotion Corpus (HaEmoC_V1). The workflow, depicted in Fig. 5, outlines the key steps involved in data collection, pre-processing, and categorization.

A. *Dataset Acquisition:* The dataset was collected using the Tweepy API, which enabled the retrieval of tweets from native Hausa speakers on Twitter. This process resulted in a corpus of 19,200 raw tweets, containing unstructured data such as text, hashtags, URLs, emoji, and other social media-specific elements. Given the noisy nature of Twitter data, systematic filtering and pre-processing were necessary to ensure the quality and relevance of the dataset.

B. *Data Preparation and Pre-processing:* To refine the dataset, multiple pre-processing steps were undertaken. Text filtering was applied by removing URLs, hashtags, numbers, punctuation marks, and non-Hausa words to ensure the dataset focused on the core textual content while maintaining the linguistic integrity of the Hausa language. Keyword-based filtering was also implemented to prioritize tweets containing specific keywords relevant to disaster relief and emergency response, such as *Tallafin Gaggawa*, *Tallafin kayan abinci*, and *Tallafin Annoba*, thereby enhancing the relevance of the dataset in line with the research objectives. In addition, tokenization and stopword removal were carried out, where the text was segmented into individual words or tokens, and common Hausa stopwords such as *da*, *ta*, and *shi* were eliminated to reduce non-informative elements and retain only meaningful words for subsequent analysis.

C. *Data Categorization and Final Dataset Construction:* After pre-processing, the dataset was further refined and reduced to 12,811 cleaned tweets. These tweets were then manually annotated and categorized into predefined emotion classes, including; *Anger, Sadness, Disgust, Fear, Surprise, Joy, Trust, Optimism, Pessimism, Anticipation, Neutral*. These categorized datasets were essential for training transformer-based models in emotion classification, ensuring a high-quality benchmark for low-resource language sentiment analysis. The entire dataset preparation framework is illustrated in Fig. 5.

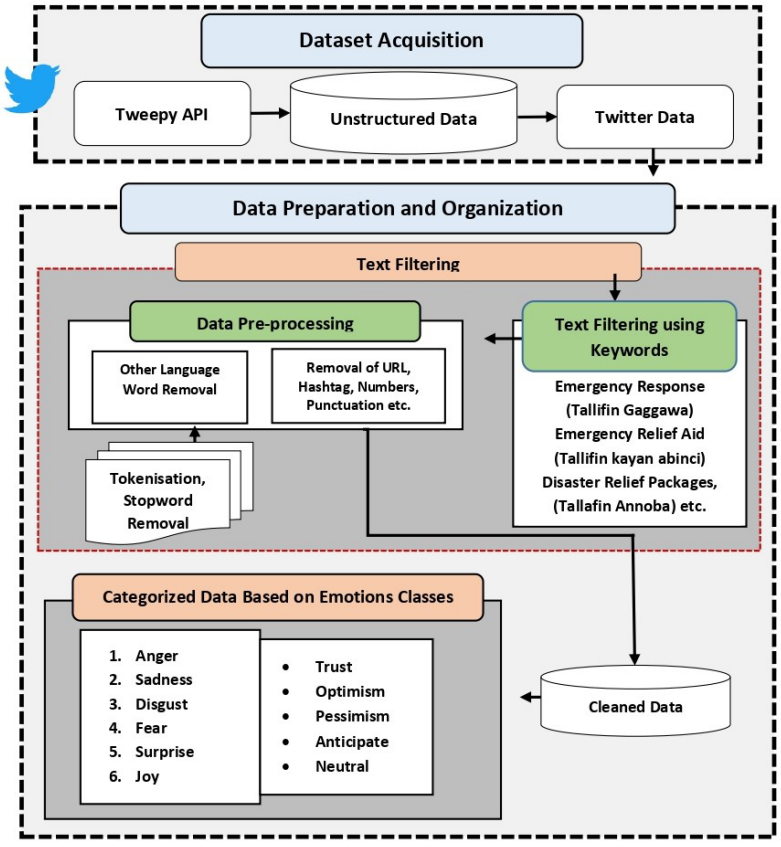

Fig. 5: Dataset acquisition framework

3.4 HaEmoC_V1 statistics

This section provides a detailed statistical overview of the Hausa Emotion Corpus (HaEmoC_V1). The dataset consists of 12,811 tweets, each labelled with one or more of 11 predefined emotion categories making the dataset suitable for multi-label emotion detection tasks. HaEmoC_V1 is structured as a textual dataset with no missing values, ensuring high data integrity for machine learning applications. Each instance includes a unique ID, a tweet text, and binary emotion labels, where 1 indicates the presence of an emotion and 0 indicates its absence. The dataset plays a key role in benchmarking transformer-based models for low-resource language processing, as shown in Table 4. For experimental evaluation, the HaEmoC_V1 dataset was divided into training (80%), validation (10%), and testing (10%) subsets using stratified multi-label sampling to preserve label distribution consistency across all partitions.

Table 4: Dataset description of HaEmoC_V1

Corpus properties	HaEmoC_V1 Description
Dataset characteristics	Textual dataset
Attribute characteristics	- Text: *Tweet content (string).* - Emotions: *11 binary attributes (0 or 1) for anger, sadness, disgust, fear, surprise, joy, trust, optimism, pessimism, anticipation, and neutral*.
Total number of instances	12,811
Missing values	No missing values
Total number of attributes	13 attributes (1 text column + 11 emotion columns + 1 ID column).

3.5 Sample of HaEmoC_V1 dataset

This section presents some samples of the instances from the HaEmoC_V1 dataset, showcasing its structure and labelling format. Each sample consists of a tweet in Hausa alongside its corresponding binary emotion labels, where 1 signifies the presence of an emotion and 0 indicates its absence as shown in Table 5.

Table 5: Sample of HaEmoC_V1

ID	Tweet	English Translation	ang	sad	dis	fea	sur	joy	tru	opt	pes	ant	neu
1	BBC idan bakuda abin posting ku saka mana hoton buhari mu zazzageshi mana	BBC if you have nothing to post, upload Buhari's photo so we can bash him	1	1	1	1	0	0	0	0	1	0	0
2	@user @user Toh fah inji 'yan magana suka ce """"ana wata ga wata""""????	Well then, as the saying goes, "while one is happening, another is coming"????	0	0	0	0	1	0	0	0	0	0	1
3	tohh allah shi taimaka	Well, may God help	0	0	0	0	0	1	1	1	0	1	0
4	@user Lolz?? ta aure shi ko ya aureta!? VOA hausa una no get news again ??! Haba since last year!	Lol?? Did she marry him or he married her!? VOA Hausa, don't you have news anymore?? Come on, since last year!	1	1	0	0	0	0	0	0	0	0	0
5	@user Ana ta diban palliatives????	They are still looting the palliatives????	0	0	1	1	0	0	0	0	0	0	0

Note. ang = Anger; sad = Sadness; dis = Disgust; fea = Fear; sur = Surprise; joy = Joy; tru = Trust; opt = Optimism; pes = Pessimism; ant = Anticipation; neu = Neutral.

These examples help illustrate the dataset's suitability for multi-label emotion classification in low-resource NLP tasks. The label cardinality of HaEmoC_V1 is 2.3, indicating that each instance contains multiple co-occurring emotions. The label density is 0.21, reflecting moderate overlap among emotion categories. These characteristics make the dataset suitable for evaluating complex multi-label learning scenarios.

3.6 Dataset verification

After completing the pre-processing stage, the initial labelling was conducted by 11 selected postgraduate students from different Universities in Northern part of Nigeria, all of whom are native Hausa speakers. Each annotator was assigned to label only one specific emotion category throughout the dataset. For example, an annotator X could only annotate anger or sadness but not multiple emotions simultaneously. To ensure consistency, the authors developed annotation guidelines to assist the annotators in maintaining accuracy during the labelling process. Following the initial annotation phase, the authors engaged three expert annotators specializing in text emotion analysis and fluent in Hausa to review and verify the initial labels. To determine the final emotion labels for each instance, a majority voting approach was employed, ensuring the reliability and objectivity of the dataset.

3.6.1 Quality of annotation

To assess the reliability of the multi-label annotations, two widely used inter-annotator agreement metrics were applied: Jaccard Coefficient and Fleiss' Kappa. The overall agreement among the three expert annotators resulted in a Jaccard Coefficient score of 84% and a Fleiss' Kappa score of 82%, indicating a high level of consistency and accuracy in the dataset annotations. These results confirm the high quality and reliability of the HaEmoC_V1 dataset for emotion classification tasks.

3.7 Dataset correlation

This section presents the correlation analysis of the HaEmoC_V1 dataset, which evaluates the relationships between different emotion labels. Fig. 6 illustrates the correlation matrix, where darker shades indicate stronger correlations between emotions. Particularly, emotions like trust and optimism (0.9) and pessimism and anger (0.8) exhibit strong positive correlations, while joy and sadness (-0.32) display negative correlations. These insights are crucial for understanding emotion co-occurrence patterns, which can improve multi-label emotion classification models.

While the correlation analysis provides insights into label co-occurrence patterns, these relationships are not inherently exploited by conventional loss functions. To address this limitation, the proposed method incorporates a KL-based correlation alignment mechanism that directly integrates the empirical co-occurrence matrix into the optimization objective, enabling explicit modelling of label dependencies.

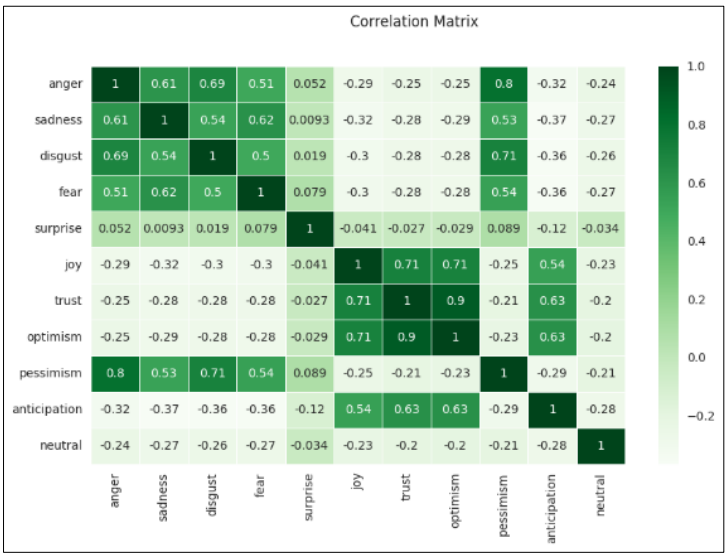

Fig. 6: Dataset correlation

3.8 Proposed loss function: Correlation-aware ConFoBCE-MLTC

This section presents the proposed Correlation-Aware Contrastive Focal Binary Cross-Entropy for Multi-Label Text Classification. The objective is to address three fundamental challenges in multi-label emotion classification: class imbalance, overlapping label distributions, and label dependency modelling. Unlike conventional loss functions that treat labels independently, the proposed approach integrates classification accuracy, representation learning, and label correlation alignment into a unified optimization framework. The overall architecture is illustrated in Fig. 7.

A. Multi-label Focal Loss (MFL): Multi-label datasets often show significant imbalance across emotion classes, which can bias model performance toward majority categories. This imbalance is also evident in the three benchmark datasets employed in this study. To address it, we adopt Multi-label Focal Loss (MFL), inspired by [92], to reduce bias and improve recognition of underrepresented emotions. MFL assigns greater weight to difficult or minority-class samples, thereby improving robustness to imbalance. The loss function is defined as in Eq. (1):

$$\mathcal{L}_{\mathrm{MFL}} = -\frac{1}{N}\sum_{i=1}^{N}\cdot\sum_{c=1}^{C}\cdot\alpha_c\left(1-p_{i,c}\right)^{\gamma}y_{i,c}\log\left(p_{i,c}\right) \tag{1}$$

where $p_{i,c}$ denotes the predicted probability, $y_{i,c}$ is the ground-truth label, α_c is the class balancing factor, and γ is the focusing parameter.

B. Multi-label Contrastive Loss with Label-Overlap Awareness (MCL): Unlike conventional contrastive loss that assumes binary similarity between samples, multi-label emotion classification requires modelling partial similarity due to overlapping labels. To address this, we define a soft similarity measure based on label overlap using the Jaccard coefficient as shown in Eq. (2).

$$s_{ij} = \frac{\left|y_i \cap y_j\right|}{\left|y_i \cup y_j\right|} \tag{2}$$

where y_i and y_j denote the label sets associated with samples *i* and *j*, respectively. This formulation enables the model to capture varying degrees of semantic similarity between samples sharing subsets of emotion labels rather than treating similarity as strictly binary.

For each mini-batch, all samples were treated as candidate contrastive pairs. Pairwise label-overlap similarity was computed using the Jaccard coefficient defined in Eq. (2). Samples exhibiting higher label overlap were assigned stronger positive similarity weights, whereas samples with minimal or no overlap implicitly contributed as negative pairs within the InfoNCE-based contrastive objective. Unlike conventional contrastive learning frameworks that rely on binary similarity assumptions, this formulation enables the model to capture varying degrees of semantic relatedness among multi-label emotion instances. Consequently, the learned embedding space becomes more discriminative and semantically structured for multi-label emotion classification.

Based on this overlap-aware similarity measure, the proposed Multi-label Contrastive Loss (MCL) is formulated as in Eq. (3):

$$L_{MCL} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}s_{ij}\log\left(\frac{\exp\left(\frac{sim(f_i,f_j)}{\tau}\right)}{\sum_{k\neq i}\exp\left(\frac{sim(f_i,f_k)}{\tau}\right)}\right) \tag{3}$$

where f_i represents the embedding of sample i, sim(·) denotes cosine similarity, and τ is a temperature parameter. This formulation enables the model to learn more discriminative representations by accounting for partial semantic similarity between multi-label samples.

C. Binary Cross-Entropy (BCE) with Logits: To maintain classification stability, we also incorporate Binary Cross-Entropy (BCE) with Logits, a standard loss widely used in multi-label classification. BCE measures the discrepancy between predicted probabilities and actual labels. Although it may struggle with imbalance

and overlapping emotions [102], it provides a strong baseline for stable training. The BCE loss is expressed as in Eq. (4):

$$\mathcal{L}_{BCE} = -\sum_{i=1}^{N} \cdot \sum_{c=1}^{C} \left[y_{i,c} \, log\left(\sigma(z_{i,c})\right) + (1 - y_{i,c}) \, log\left(1 - \sigma(z_{i,c})\right) \right] \tag{4}$$

where σ(·) is the sigmoid function.

D. Label Correlation Alignment (KL-based): To explicitly model dependencies among emotion labels, a label correlation alignment component based on Kullback–Leibler (KL) divergence is introduced. Let $A \in \mathbb{R}^{C \times C}$ denote the empirical label co-occurrence matrix derived from the ground-truth labels, and $Q \in \mathbb{R}^{C \times C}$ represent the predicted co-occurrence matrix computed from the model outputs. These matrices are formulated in Eq. (5):

$$Q = \hat{Y}^T \hat{Y}, \qquad A = Y^T Y \tag{5}$$

where Y denotes the ground-truth label matrix and $\hat{Y}$ represents the predicted label probability matrix. The empirical and predicted co-occurrence matrices were normalized row-wise to form valid probability distributions prior to KL divergence computation. The normalized matrices are computed in Eq. (6):

$$\tilde{A}_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}$$
$$\tilde{Q}_{ij} = \frac{Q_{ij}}{\sum_j Q_{ij}} \tag{6}$$

To ensure numerical stability and avoid undefined logarithmic operations, a small constant $\varepsilon = 10^{-8}$ was added to both matrices prior to KL divergence computation. After normalization, the KL-based label correlation alignment loss is defined in Eq. (7):

$$L_{\mathrm{KL}} = \sum_{i=1}^{C} \sum_{j=1}^{C} \tilde{A}_{ij} \log\left(\frac{\tilde{A}_{ij}}{\tilde{Q}_{ij}}\right) \tag{7}$$

This component encourages consistency between the predicted and empirically observed label relationships by aligning the learned co-occurrence structure with the ground-truth label dependency distribution. Consequently, the model is better able to capture inter-label semantic dependencies, leading to more coherent and contextually consistent multi-label predictions.

E. Final Loss (ConFoBCE-MLTC): The final ConFoBCE-MLTC loss integrates the complementary strengths of MFL, MCL, and BCE. This hybridization addresses imbalance by focusing on minority emotions, enhances emotion separability through contrastive learning, and ensures overall training stability with BCE. Additionally, a KL divergence term (L_{KL}) is incorporated to regularize the learned distributions. The combined loss is defined in Eq. (8):

$$L_{\_total} = L_{BCE} + \lambda\_f \, . \, L_{MFL} + \lambda\_c \, . \, L_{MCL} + \lambda\_k \, . \, L_{KL} \tag{8}$$

where λ_f, λ_c, λ_k control the contributions of focal loss, contrastive learning, and correlation alignment respectively.

3.8.1 Architectural Framework and Training Pipeline

This section presents the architectural framework and end-to-end training pipeline of the proposed ConFoBCE-MLTC model for transformer-based multi-label emotion classification. The framework is designed to jointly optimize classification performance, discriminative feature representation, and label dependency modelling by integrating multiple complementary learning objectives within a unified optimization strategy. Fig. 7 and Algorithm 1 illustrate the overall system architecture, highlighting the flow from transformer-based input encoding through multi-stream learning, unified loss computation, and gradient-based parameter optimization.

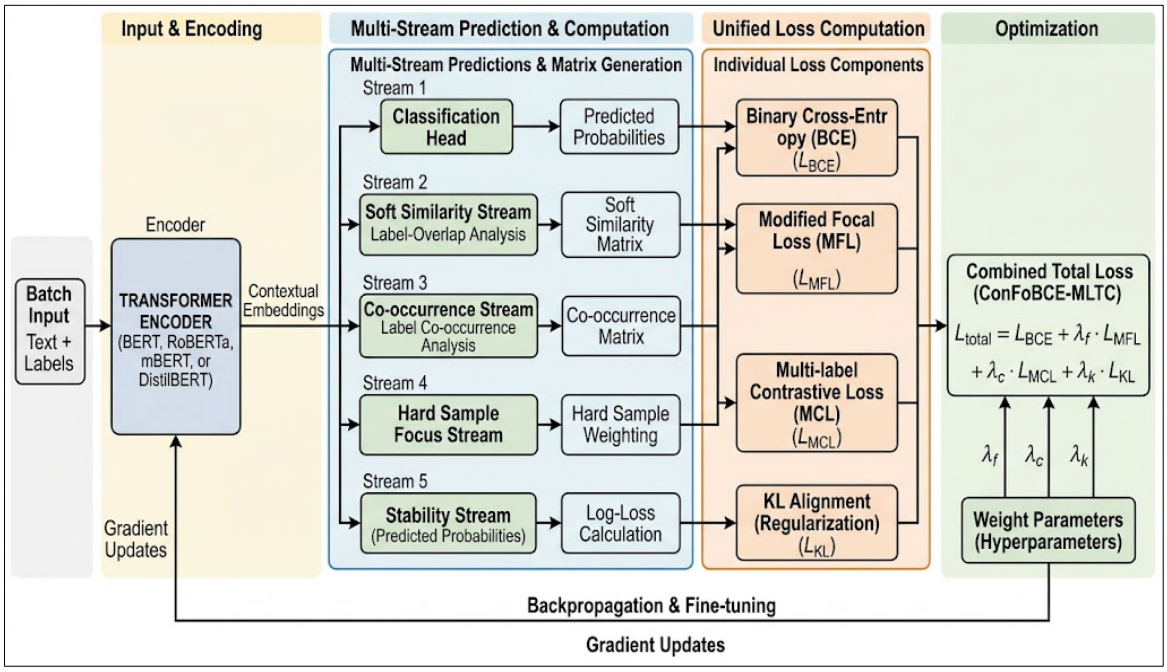

Fig. 7: End-to-end training architecture of the proposed ConFoBCE-MLTC loss framework for transformer-based multi-label emotion classification.

The architectural components and training workflow illustrated in Fig. 7 demonstrate the complete end-to-end implementation of the proposed ConFoBCE-MLTC framework. The process begins with a mini-batch of input text-label pairs, which are encoded using a transformer backbone (BERT, RoBERTa, mBERT, or DistilBERT) to generate contextual embeddings. These embeddings are subsequently processed through five complementary streams: the prediction stream for label probability estimation, the soft similarity stream for label-overlap based contrastive learning, the co-occurrence stream for empirical label dependency modelling, the hard sample focus stream for emphasizing difficult samples, and the stability stream for ensuring prediction consistency during training.

The outputs from these streams are integrated through a unified loss computation module, where Binary Cross-Entropy (BCE) supports baseline classification, Multi-label Focal Loss (MFL) addresses class imbalance and hard-example learning, Multi-label Contrastive Loss (MCL) improves embedding discrimination using soft label similarity, and KL Alignment Loss preserves empirical label correlation structure. These four objectives are combined into the proposed ConFoBCE-MLTC loss using weighted hyperparameters $\lambda_f, \lambda_c, \lambda_k$ after which the total loss is backpropagated to fine-tune the transformer encoder in an end-to-end manner. This unified optimization framework enables simultaneous learning of classification accuracy, robust semantic representation, and inter-label dependency awareness, thereby improving generalization performance, particularly in low-resource multilingual emotion classification settings. The overall architectural framework and training procedure is also presented in Algorithm 1.

Algorithm 1: *Modelling with the Proposed Loss Function (ConFoBCE-MLTC)*

Input:
  D ← input_dataset
  D_tr, D_val ← split(D)
Initialize:
  θ ← initialize_model_parameters
  αf, γf ← focal_loss_parameters
  λ_f, λ_c, λ_k ← loss_weights
  ε ← 10^(-8)
for epoch ∈ (1, E_max) do
  for each batch B ∈ D_tr do
    X, Y ← extract_inputs_and_labels(B)
    // **Step 1: Forward propagation**
    Z_logits ← fθ(X)
    Ŷ ← sigmoid(Z_logits)
    // **Step 2: Classification losses**
    L_BCE ← BCEWithLogits(Z_logits, Y)
    L_MFL ← FocalLoss(Ŷ, Y; αf, γf)
    // **Step 3: Label correlation matrices**
    Q ← ŶᵀŶ
    A ← YᵀY
    // **Step 4: Row-wise normalization**
    Q ← normalize_rows(Q + ε)
    A ← normalize_rows(A + ε)
    // **Step 5: KL-based label correlation alignment**
    L_KL ← KL_divergence(A, Q)
    // **Step 6: Contrastive learning branch**
    if contrastive_pairs_available(B) then
      Z ← extract_embeddings(B)
      // **Step 6.1: Compute Jaccard label-overlap similarity**
      S ← compute_label_overlap_matrix(Y)
      // **Step 6.2: Multi-label contrastive loss**
      L_MCL ← multi_label_contrastive(Z, S)
      // **Step 6.3: Final combined loss**
      L_total ← L_BCE
          + λ_f · L_MFL
          + λ_c · L_MCL
          + λ_k · L_KL
    else
      // **Step 6.4: Loss without contrastive component**
      L_total ← L_BCE
          + λ_f · L_MFL
          + λ_k · L_KL
    end if
    // **Step 7: Backpropagation and parameter update**
    θ ← θ − η · ∇θ(L_total)
  end for
end for
// **Step 8: Model evaluation**
Eval ← evaluate_model(fθ, D_val)

4.0 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental Setup and Evaluation Metrics

The effectiveness of the proposed ConFoBCE-MLTC loss function was evaluated using four transformer-based architectures: BERT, DistilBERT, RoBERTa, and mBERT. Each model was fine-tuned on three multi-label datasets representing both high-resource and low-resource settings: SemEval-2018 Task 1 (English), Bahasa Indonesian Hate Speech (BHIS), and the proposed Hausa Emotion Corpus (HaEmoC_V1).

All experiments were conducted on an NVIDIA RTX 4090 GPU with 24 GB VRAM using PyTorch and HuggingFace Transformers. To ensure a fair and reproducible comparison, all models were trained under identical experimental conditions. Input sequences were tokenized using model-specific tokenizers and standardized to a maximum sequence length of 128 tokens through truncation and padding. Early stopping based on validation loss was further employed to mitigate overfitting during training. Table 6 summarizes the detailed experimental and training configurations adopted across all experiments.

Table 6: Experimental and Training Configuration for ConFoBCE-MLTC

Parameter	Configuration
GPU	NVIDIA RTX 4090 (24 GB VRAM)
Framework	PyTorch + HuggingFace Transformers
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	16
Epochs	30
Weight Decay	0.01
Warmup Ratio	0.1
Dropout Rate	0.3
Random Seed	42
Sigmoid Threshold	0.5
Temperature Parameter (τ)	0.07
Focal Gamma (γ)	2.0
Alpha Balancing Strategy	Inverse Class Frequency

Performance was evaluated using Micro-F1, Macro-F1, and Jaccard Score, which together capture overall classification effectiveness, class-level balance, and label overlap consistency in multi-label settings. Micro-F1 emphasizes global predictive performance across all labels, whereas Macro-F1 provides a balanced assessment across both frequent and underrepresented classes. Jaccard Score further measures the similarity between predicted and ground-truth label sets, making it particularly suitable for evaluating multi-label emotion classification tasks.

4.2 Results Analysis and Discussion

This section provides a comprehensive evaluation of the proposed approach. The experiments systematically assess the effectiveness of various transformer architectures under multiple loss function configurations, highlighting their impact on multi-label emotion classification performance. In addition, hyperparameter tuning was performed to determine optimal values for the weighting coefficients of the proposed loss function. A sensitivity analysis was conducted to evaluate the robustness of these parameters across different configurations.

4.2.1 Baseline Performance Across Transformer Models

Table 7 reports the baseline performance of all transformer models evaluated using three widely adopted loss functions: Binary Cross-Entropy (BCE), Focal Loss, and Contrastive Loss. These results establish a reference point for comparing the effectiveness of the proposed method against conventional training objectives.

Table 7: Baseline loss weighing performance across Transformer Models

Model	Loss Function	Dataset	Micro-F1	Macro-F1	Jaccard Score
BERT	BCE	HaEmoC_V1	74.18	71.22	61.18
		BHIS	79.31	77.14	67.42
		SE-18-T1 E-c	80.26	78.21	69.33
	Focal	HaEmoC_V1	76.33	73.28	64.12
		BHIS	77.21	75.16	66.28
		SE-18-T1 E-c	81.27	79.33	70.22
	Contrastive	HaEmoC_V1	77.42	74.31	65.31
		BHIS	78.36	76.22	67.36
		SE-18-T1 E-c	82.14	80.27	71.44
DistilBERT	BCE	HaEmoC_V1	76.22	73.19	63.22
		BHIS	79.16	77.28	68.41
		SE-18-T1 E-c	81.28	79.24	70.36
	Focal	HaEmoC_V1	76.11	73.08	64.29
		BHIS	79.23	76.32	67.18
		SE-18-T1 E-c	80.14	78.17	69.21
	Contrastive	HaEmoC_V1	78.37	75.29	66.35
		BHIS	80.24	78.19	69.48
		SE-18-T1 E-c	82.31	80.22	71.26
RoBERTa	BCE	HaEmoC_V1	68.21	65.18	57.11
		BHIS	74.27	72.21	63.27
		SE-18-T1 E-c	80.19	78.16	69.41
	Focal	HaEmoC_V1	70.33	67.26	59.28
		BHIS	76.18	74.22	65.12
		SE-18-T1 E-c	81.14	79.27	70.37
	Contrastive	HaEmoC_V1	72.24	69.18	61.17
		BHIS	78.11	76.19	67.42
		SE-18-T1 E-c	81.21	79.16	71.33
mBERT	BCE	HaEmoC_V1	75.28	73.14	64.18
		BHIS	80.22	78.18	69.31
		SE-18-T1 E-c	83.31	81.24	72.14
	Focal	HaEmoC_V1	77.33	75.21	66.27
		BHIS	82.18	80.26	71.39
		SE-18-T1 E-c	84.27	82.22	73.21
	Contrastive	HaEmoC_V1	**79.14**	77.06	68.33
		BHIS	**82.24**	80.18	71.22
		SE-18-T1 E-c	**83.17**	81.21	72.27

A consistent trend is observed across all models and datasets as illustrated in Table 7 above. performance improves progressively from BCE to Focal Loss and further to Contrastive Loss. This trend highlights two important observations. First, BCE, while stable, is insufficient for handling the complexities of multi-label emotion classification, particularly in the presence of class imbalance and label overlap. Second, the inclusion of focal weighting improves sensitivity to minority classes, leading to moderate gains in Macro-F1. However, the most significant improvements arise from contrastive learning, which enhances representation separability by structuring the embedding space.

Among the evaluated models, mBERT consistently achieves the strongest baseline performance, particularly on the low-resource Hausa dataset (HaEmoC_V1), where it reaches 79.14 Micro-F1 using Contrastive Loss. This result underscores the advantage of multilingual pretraining, which provides a strong inductive bias for cross-lingual generalization. In contrast, RoBERTa exhibits relatively weaker performance on HaEmoC_V1, suggesting limited transferability of monolingual representations to low-resource and morphologically diverse languages. However, its performance improves on high-resource datasets, indicating that its effectiveness remains domain-dependent.

Interestingly, DistilBERT demonstrates competitive performance despite its reduced parameter size, particularly under contrastive learning. This suggests that representation quality, rather than model size alone, is a key determinant of performance in multi-label classification tasks.

4.2.2 Performance of the Proposed ConFoBCE-MLTC Loss

The performance of the proposed ConFoBCE-MLTC loss function is presented in Table 8. Across all models and datasets, the proposed approach consistently outperforms baseline loss functions, confirming the effectiveness of integrating focal modulation, contrastive learning, and label correlation alignment within a unified optimization framework.

Table 8: Performance of the proposed loss function

Proposed Loss	Model +	Dataset	Micro-F1	Macro-F1	Jaccard Score
ConFoBCE-MLTC	BERT	HaEmoC_V1	78.31	76.28	68.26
		BHIS	83.18	81.22	73.18
		SE-18-T1 E-c	84.22	82.19	74.21
	DistilBERT	HaEmoC_V1	79.26	77.21	69.17
		BHIS	81.21	79.28	71.22
		SE-18-T1 E-c	83.27	81.24	73.19
	RoBERTa	HaEmoC_V1	73.18	70.26	63.21
		BHIS	78.12	76.21	68.36
		SE-18-T1 E-c	83.19	81.25	73.22
	mBERT	HaEmoC_V1	**82.24**	80.18	73.31
		BHIS	**85.19**	83.23	75.22
		SE-18-T1 E-c	**87.18**	85.21	78.36

Based on the results presented in Table 8, the most significant gains are observed with mBERT + ConFoBCE-MLTC, which achieves strong empirical performance across all datasets, including 82.24 Micro-F1 on HaEmoC_V1, 85.19 on BHIS, and 87.18 on SemEval-2018. These results highlight the strength of combining focal loss, contrastive learning, and label correlation alignment in a unified framework.

Compared to baseline contrastive learning, the proposed method improves performance by approximately 3–5% in Micro-F1 and up to 6% in Jaccard Score, particularly in low-resource settings. This demonstrates that incorporating label dependency modeling significantly enhances multi-label prediction consistency. However, even weaker baseline models such as RoBERTa benefit substantially from the proposed loss, achieving up to 73.22 Jaccard Score on SemEval-2018. This indicates that the improvement is not architecture-dependent but rather driven by the effectiveness of the loss design.

To further strengthen the evaluation, we compare the proposed method with Asymmetric Loss (ASL) as presented in Table 9 [102], a recent state-of-the-art approach for multi-label classification. ASL was selected because it remains the most widely adopted and reproducible asymmetric multi-label loss baseline in recent literature. The results indicate that ConFoBCE-MLTC consistently outperforms ASL across all metrics as shown in Table 8. The improvement in Jaccard Score highlights the effectiveness of the correlation-aware component in modelling label co-occurrence structures beyond asymmetric weighting strategies.

Table 9: Performance of Asymmetric loss

Loss Function	Model	Dataset	Micro-F1	Macro-F1	Jaccard Score
Asymmetric Loss	BERT	HaEmoC_V1	75.41	72.35	63.18
		BHIS	80.22	78.14	68.27
		SE-18-T1 E-c	81.33	79.28	70.19
	DistilBERT	HaEmoC_V1	77.18	74.22	64.31
		BHIS	80.31	78.19	69.14
		SE-18-T1 E-c	81.29	79.18	70.24
	RoBERTa	HaEmoC_V1	69.28	66.21	58.33
		BHIS	75.19	73.14	64.28
		SE-18-T1 E-c	80.33	78.22	69.47
	mBERT	HaEmoC_V1	**78.22**	75.31	67.24
		BHIS	**83.17**	81.19	72.33
		SE-18-T1 E-c	**85.21**	83.18	74.29

As presented in Table 9 above. The Asymmetric Loss consistently improves over standard BCE across all models and datasets, yielding Micro-F1 gains of approximately 1–3%. However, it remains inferior to the proposed ConFoBCE-MLTC framework. mBERT + Asymmetric Loss achieves the strongest results (85.21 Micro-F1 on

SemEval-2018), reinforcing the advantage of multilingual pretraining. DistilBERT remains competitive despite its smaller size, while RoBERTa continues to struggle on low-resource HaEmoC_V1 (69.28 Micro-F1) due to limited cross-lingual transferability. Overall, Asymmetric Loss provides meaningful but incremental improvements over BCE and Focal by addressing false positive-negative imbalance. Nevertheless, it lacks the contrastive learning and label correlation alignment components essential for achieving strong empirical performance, confirming the superiority of the proposed ConFoBCE-MLTC approach.

4.2.3 Confusion Matrix and Class-wise Analysis

To investigate the behaviour of the best-performing model, a detailed class-wise analysis was conducted using both the classification report presented in Fig. 8 and the confusion matrix illustrated in Fig. 9. While Table 2 demonstrates strong aggregate performance (Micro-F1 = 87.18, Macro-F1 = 85.21), these metrics do not fully capture how performance is distributed across individual emotion categories. The class-wise evaluation reveals consistently strong performance across most labels, with F1-scores exceeding 0.85 for the majority of emotions. Notably, trust and optimism achieve the highest performance (F1 = 0.90), followed closely by anger, sadness, and joy, suggesting that emotions with clearer contextual signals benefit significantly from the combined effects of focal modulation and contrastive representation learning.

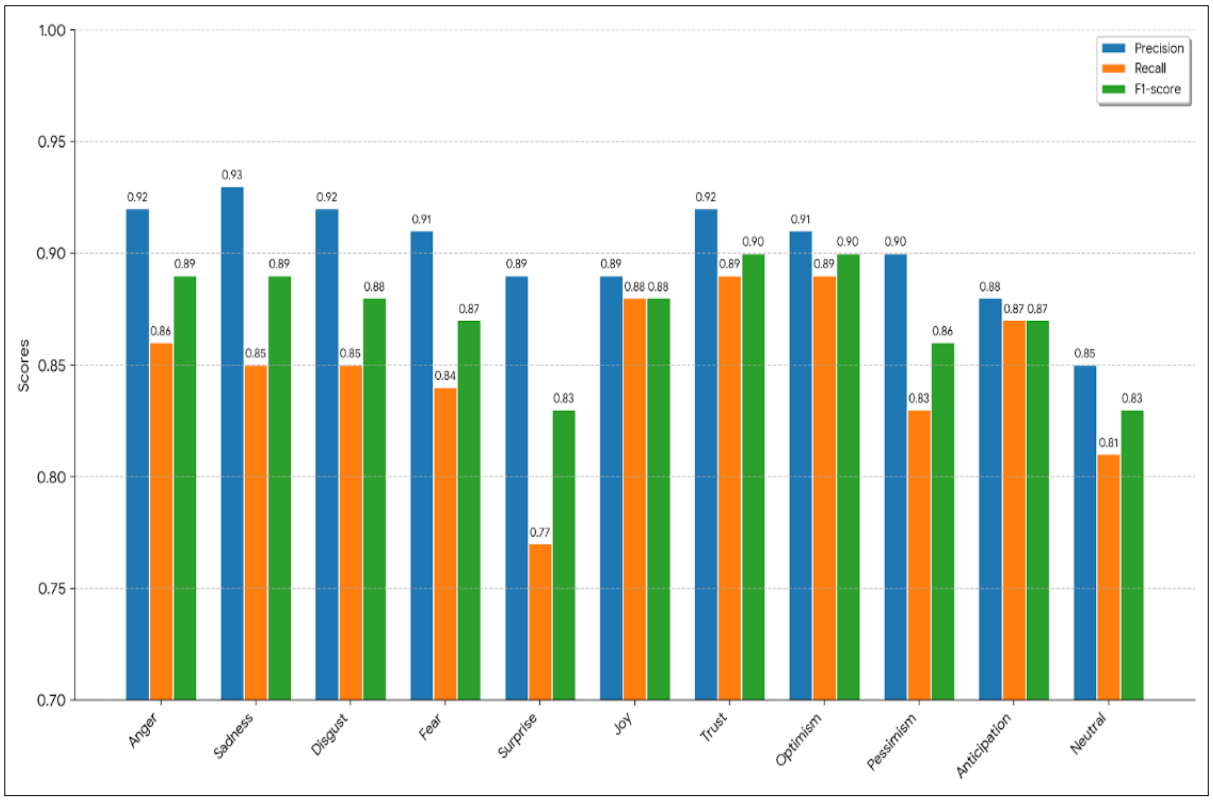

Fig. 8: Class-wise Performance of mBERT + ConFoBCE-MLTC

Despite these strong results, the class-wise analysis also highlights limitations. Lower performance is observed for surprise (F1 = 0.83) and neutral (F1 = 0.83), indicating that these categories remain challenging due to their inherent ambiguity and overlap with other emotions. A consistent gap between precision and recall is also evident, with precision generally higher, suggesting a conservative prediction strategy that prioritizes correctness over coverage. While this reduces false positives, it leads to under-prediction of certain labels, particularly in borderline or overlapping cases.

This behaviour is further illustrated by the confusion matrix presented in Fig. 9, which provides deeper insight into model prediction patterns. A strong diagonal dominance is observed, confirming high correct classification rates across most emotion categories, particularly for sadness, trust, optimism, anticipation, and joy. However, the matrix also reveals structured misclassification patterns rather than random errors, with confusion concentrated among semantically related emotions. For instance, surprise is frequently misclassified as fear and joy, while pessimism overlaps with sadness and fear, and neutral is distributed across multiple categories due to weak emotional signals.

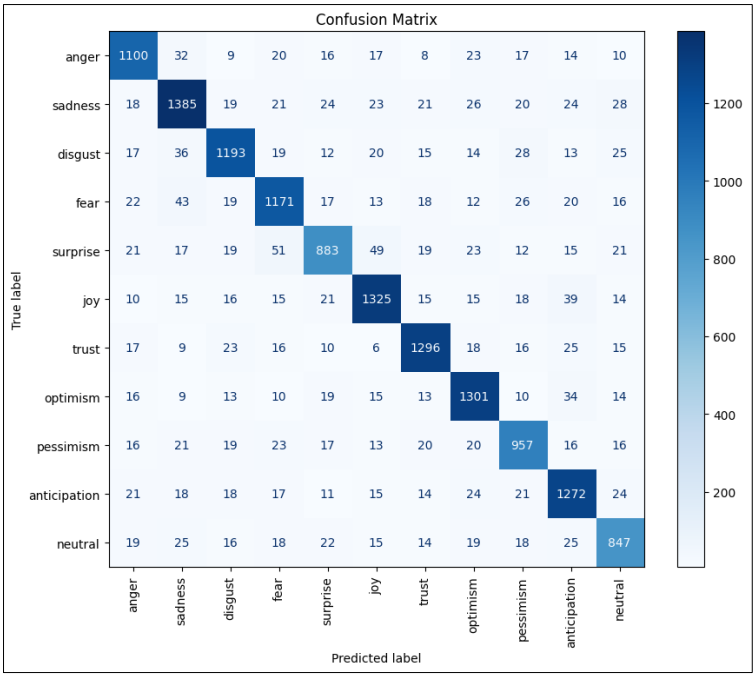

Fig. 9: Confusion Matrix of mBERT + ConFoBCE-MLTC on HaEmoC_V1

Furthermore, fine-grained overlaps among positive emotions, such as between optimism and anticipation, and between trust and joy, highlight the challenge of distinguishing fine-grained emotional distinctions. Importantly, these errors are not arbitrary but reflect meaningful semantic relationships, indicating that the model has learned structured representations consistent with the design of the proposed loss function. Overall, the combined evidence from the classification report and confusion matrix corroborates the improvements reported in Table 8, demonstrating that performance gains are both globally significant and locally consistent across classes. At the same time, the remaining errors emphasize the inherent complexity of multi-label emotion classification, suggesting that future work should focus on modelling deeper contextual and higher-order label dependencies.

4.2.4 Hyperparameter Sensitivity Analysis

To evaluate the robustness of the proposed ConFoBCE-MLTC loss function, we conducted a sensitivity analysis on the weighting coefficients λ_f (focal loss), λ_c (contrastive loss), and λ_k (correlation alignment) as presented in Fig. 10. These parameters control the contribution of each component to the overall objective function. We varied each parameter independently while keeping others fixed and evaluated performance on the HaEmoC_V1 dataset using mBERT.

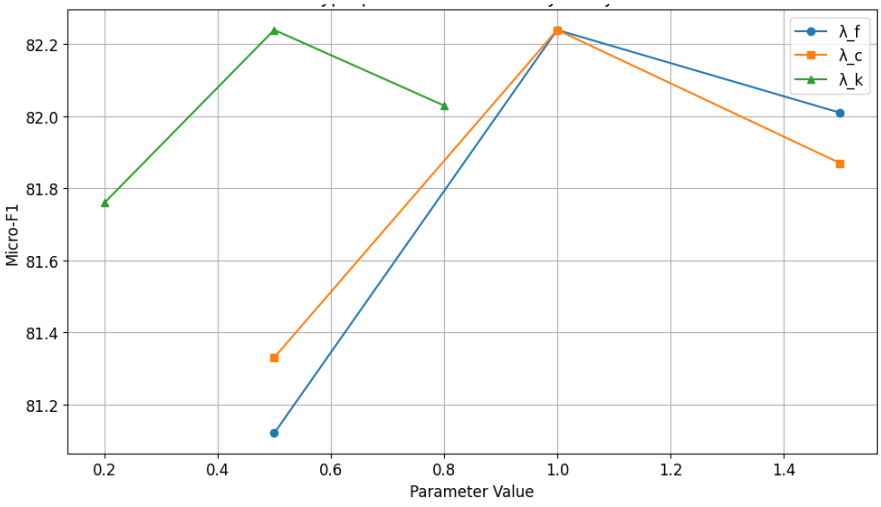

Fig. 10: Sensitivity analysis of λ parameters (mBERT on HaEmoC_V1)

The hyperparameter sensitivity analysis presented in Fig. 10 highlights the balanced contribution of each loss component toward the overall effectiveness of the proposed framework. Variations in the focal loss coefficient (λ_f) reveal that increasing its contribution generally enhances Macro-F1 performance, demonstrating improved recognition of underrepresented emotional categories. The best Macro-F1 score is achieved around λ_f = 1.5, suggesting that moderate emphasis on difficult samples strengthens minority-class discrimination without significantly compromising overall stability. However, the marginal fluctuation in Micro-F1 at higher λ_f values indicates that excessive weighting may slightly bias optimization toward hard instances.

Adjustments to the contrastive loss weight (λ_c) further demonstrate the importance of representation learning in multi-label emotion classification. Performance consistently improves as λ_c increases from 0.5 to 1.0, where the highest overall balance between Micro-F1, Macro-F1, and Jaccard Score is obtained. This behavior suggests that contrastive supervision enhances semantic separation among emotional representations, enabling the encoder to better capture inter-label distinctions. Nevertheless, performance slightly declines when λ_c exceeds the optimal range, implying that overly strong clustering constraints may reduce representation flexibility.

The KL-divergence alignment coefficient (λ_k) shows a particularly noticeable influence on Jaccard Score, emphasizing its effectiveness in preserving label dependency structures. Increasing λ_k from 0.2 to 0.5 produces measurable gains across all evaluation metrics, confirming that correlation-aware optimization improves multi-label consistency. Although performance remains competitive at λ_k = 0.8, the slight reduction compared with λ_k = 0.5 suggests that excessive regularization may constrain the model's adaptability to complex emotional distributions.

Overall, the results demonstrate that the proposed hybrid objective maintains strong and consistent performance across diverse parameter configurations. The relatively small metric fluctuations indicate that the framework is not highly sensitive to hyperparameter selection, which strengthens its practical applicability and reproducibility across different datasets and experimental settings.

4.2.5 Training and validation performance

The training and validation curves, illustrated in Fig. 11, provide insights into the learning dynamics of the best-performing model (mBERT + ConFoBCE-MLTC).

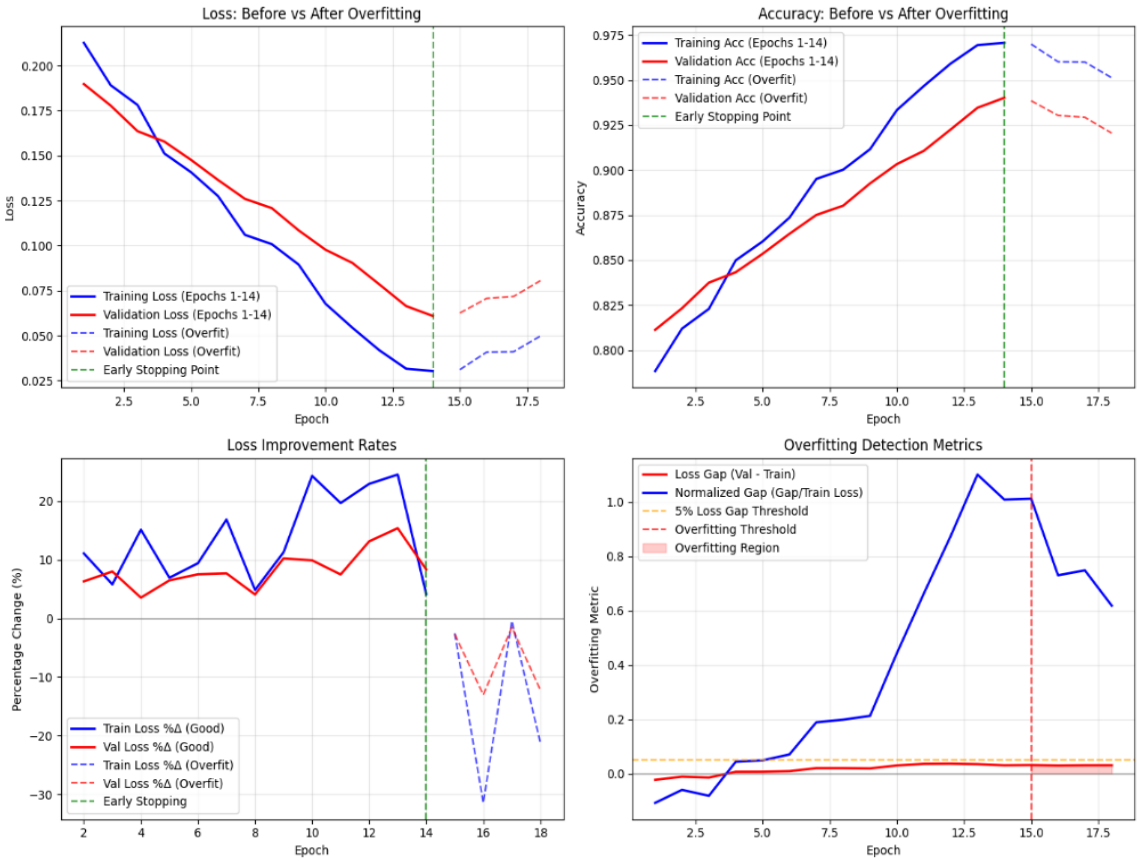

Fig. 11: mBERT Models Training & Validation Curve

During the initial training phase (epochs 1–14), both training and validation losses decrease steadily, accompanied by improvements in accuracy, indicating effective learning and good generalization. Beyond epoch 14, a divergence between training and validation performance becomes evident. While training loss continues to decrease, validation loss begins to increase, signaling the onset of overfitting. This behavior is further confirmed by a widening loss gap and declining validation accuracy. These observations justify the use of early stopping at epoch 14, which corresponds to the optimal trade-off between learning capacity and generalization performance. Importantly, the stability observed during early training suggests that the proposed loss function does not introduce optimization instability, despite its composite nature.

4.2.6 Computational Complexity and Efficiency

While the proposed ConFoBCE-MLTC demonstrates superior predictive performance, it introduces additional computational overhead due to the pairwise similarity computations required by its contrastive regularization component. Unlike conventional loss functions such as Binary Cross-Entropy (BCE) and Focal Loss, which operate with linear batch complexity, the contrastive component performs pairwise sample interactions within each mini-batch, increasing the training complexity from O(N) to $O(N^2)$, where *N* denotes the batch size.

To assess the practical implications of this design choice, computational efficiency was evaluated using three criteria: average training time per epoch, GPU memory utilization, and theoretical computational complexity, as summarized in Table 10.

Table 10: Computational Complexity and Efficiency Analysis

Loss Function	Time/Epoch	GPU Memory	Complexity
BCE	1.0×	100%	O(N)
Focal Loss	1.1×	105%	O(N)
Contrastive Loss	1.4×	120%	$O(N^2)$
ConFoBCE-MLTC	1.5×	125%	$O(N^2)$

As shown in Table 10, ConFoBCE-MLTC incurs approximately 50% additional training time and 25% higher GPU memory usage compared to BCE, primarily due to the construction of the pairwise similarity matrix required for contrastive optimization. However, this overhead remains computationally manageable because the quadratic operation is restricted to the training phase and benefits from efficient mini-batch optimization on modern GPU architectures. Importantly, this additional cost does not significantly affect inference efficiency, as prediction relies only on the transformer forward pass, maintaining near-linear complexity.

Therefore, the observed computational increase represents a deliberate and acceptable performance efficiency trade-off, where moderate training overhead yields substantial and consistent improvements in Micro-F1, Macro-F1, and Jaccard Index across all benchmark datasets. This demonstrates that the proposed method achieves enhanced predictive robustness while preserving practical deployability for real-world multi-label emotion classification tasks.

To further illustrate the computational trade-off introduced by the proposed method, Fig. 12 presents a comparative visualization of training time and GPU memory consumption across all evaluated loss functions, highlighting the moderate but manageable overhead of ConFoBCE-MLTC.

4.2.7 Ablation Study of the Best Performing Model (mBERT)

To quantify the contribution of each component of the proposed loss function, an ablation study was conducted using mBERT as the backbone model. The results are summarized in Table 11, while the comparison is illustrated in Fig. 13.

The ablation results reveal a clear and progressive improvement when additional components are integrated into the loss function. Starting from BCE, performance steadily increases with Focal Loss and further with Contrastive Loss, indicating that both imbalance handling and representation learning are crucial for multi-label emotion classification. The combination of BCE and Focal Loss yields moderate improvements, while integrating BCE with Contrastive Loss provides stronger gains, particularly in Jaccard Score, suggesting improved label overlap handling. However, the most significant improvement is observed when both Focal and Contrastive losses are combined, confirming their complementary roles.

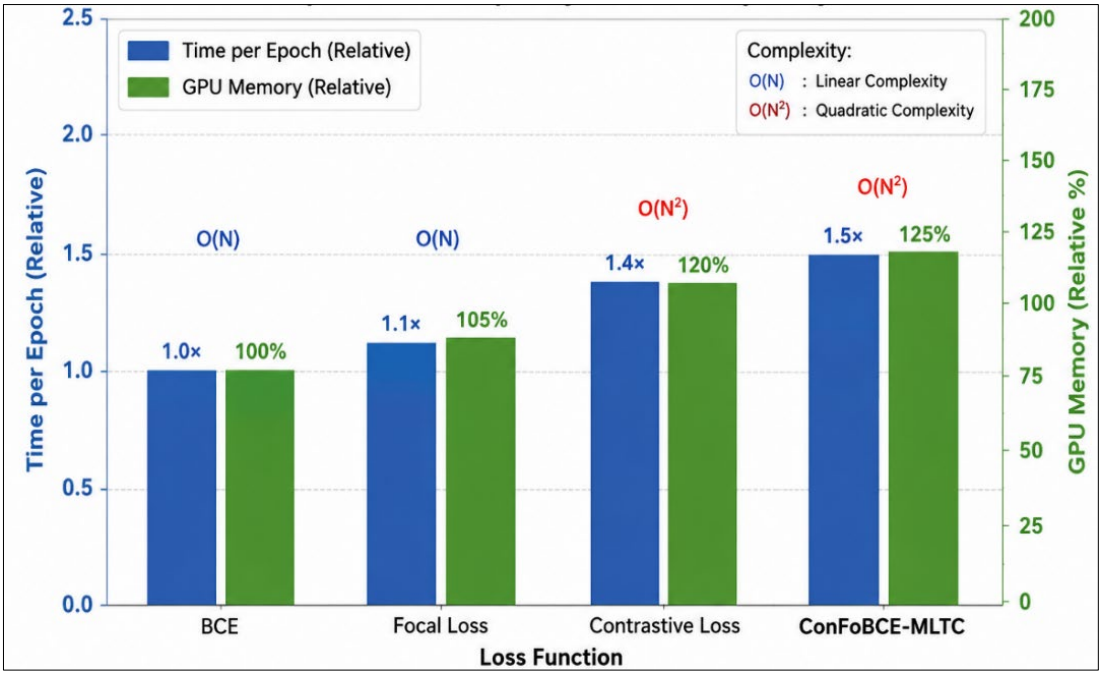

Fig. 12: Performance efficiency trade-off illustrating the relationship between computational cost and predictive performance of different loss functions.

Table 11: Ablation study of the best performing model mBERT

Loss functions configurations	HaEmoC_V1 dataset			BHIS dataset			SE-18-T1 E-c		
	Micro-F1	Macro-F1	Jaccard Score	Micro-F1	Macro-F1	Jaccard Score	Micro-F1	Macro-F1	Jaccard Score
BCE + Focal	78.62	76.48	67.52	80.73	79.64	72.18	81.58	80.63	73.86
BCE + Contrastive	79.68	77.59	69.21	81.11	80.02	72.64	83.16	81.08	74.52
Focal + Contrastive	80.92	78.69	70.74	83.36	81.27	73.11	84.43	83.36	75.03
BCE + Focal + Contrastive	81.32	79.17	72.45	84.41	82.52	74.11	85.12	83.76	76.25
+ KL (Full)	**82.24**	**80.18**	**73.31**	**85.19**	**83.23**	**75.22**	**87.18**	**85.21**	**78.36**

Note: HaEmoC_V1 = Hausa Emotion Corpus; BHIS = Bahasa Indonesian Hate Speech dataset; SE-18-T1 E-c = SemEval-2018 Task 1 E-c, Micro-F1: Micro-averaged F1-score; ConFoBCE = Contrastive Focal Binary Cross-Entropy (BCE + Focal + Contrastive); KL = Kullback–Leibler.

Finally, the full ConFoBCE-MLTC formulation achieves the best performance across all datasets, with 82.24 Micro-F1 on HaEmoC_V1, 85.19 on BHIS, and 87.18 on SemEval-2018. This represents a substantial improvement over all baseline configurations. The inclusion of the KL-based correlation alignment yields the most significant improvement in Jaccard Score, confirming that explicit modelling of label co-occurrence enhances prediction consistency. This validates the importance of incorporating structured label relationships in multi-label learning.

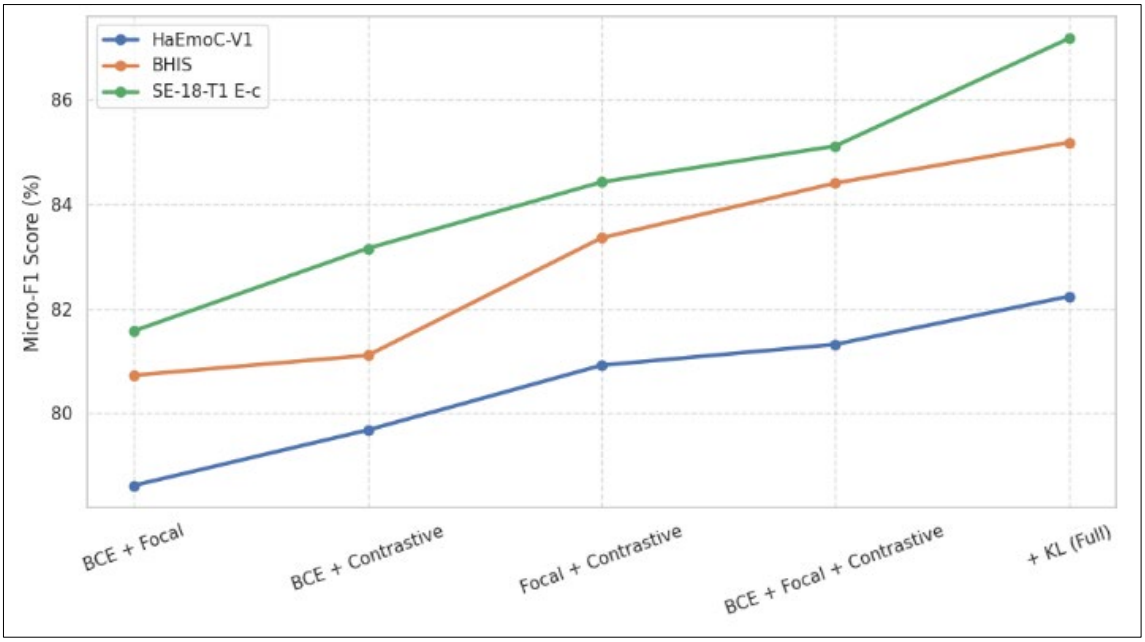

Fig. 13: Ablation comparison

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study introduced ConFoBCE-MLTC, a correlation-aware hybrid loss function designed to address fundamental challenges in multi-label emotion classification, including class imbalance, overlapping label distributions, and inter-label dependencies. By integrating focal modulation, contrastive representation learning, and probabilistic classification within a unified framework, the proposed approach enables more robust and discriminative learning without requiring modifications to underlying model architectures.

Extensive experiments across multiple datasets and transformer architectures demonstrate that ConFoBCE-MLTC consistently outperforms conventional loss functions, with the most significant improvements observed in Macro-F1 and Jaccard Score indicating improved performance on minority classes and more effective modelling of label co-occurrence structures. The superior performance of mBERT further highlights the importance of multilingual pretraining in low-resource settings. Importantly, the results confirm that explicit modeling of label correlations is a critical factor in advancing multi-label classification performance, particularly in domains characterized by high semantic overlap, such as emotion detection.

In addition, the proposed method demonstrates robustness across different hyperparameter settings and Minor improvement only between computational complexity and performance. These findings highlight its potential for practical deployment in real-world multi-label classification tasks.

5.2 Future Work

Despite the strong empirical performance achieved by the proposed framework, several limitations remain. First, the integration of pairwise contrastive operations introduces additional computational overhead during training, which may affect scalability for larger datasets and real-time applications. Second, the experimental evaluation was limited to three datasets, and broader validation across additional multilingual and low-resource benchmarks is necessary to further assess the generalizability of the proposed approach.

Future research can extend this work in several important directions. Graph-based architectures, particularly Graph Neural Networks (GNNs), could be explored to capture higher-order label dependencies beyond pairwise correlations. In addition, integrating large language models (LLMs) may further enhance contextual understanding and improve the detection of implicit emotional expressions. Cross-lingual transfer learning frameworks also present a promising direction for improving generalization across underrepresented African and low-resource languages. Finally, developing lightweight and computationally efficient variants of the proposed ConFoBCE-MLTC framework remains essential for facilitating practical deployment in resource-constrained environments and real-world multilingual applications.

Acknowledgements
The first author gratefully acknowledges the support of the Petroleum Technology Development Fund (PTDF) for sponsoring his PhD study and Universiti Putra Malaysia (UPM) for funding the publication of this journal.

Author contributions
Hassan Adamu: Conceptualization, Investigation, Data curation, Methodology, Formal analysis, Writing – original draft, **Masrah Azrifah Azmi Murad:** Conceptualization, Validation, Writing – review & editing**,** Supervision. **Nurul Amelina Nasharuddin:** Conceptualization**,** Writing – review & editing, Supervision.

Funding
This research was funded by Universiti Putra Malaysia (UPM) under the Geran Putra Inisiatif (GPI/2025/983400).

Data availability
The data that support the findings of this study are available from the corresponding author Hassan Adamu (gs66245@student.upm.edu.my) upon reasonable request.

Declaration of competing interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] A. A. L. Maruf, F. Khanam, and M. Haque, “Challenges and Opportunities of Text-Based Emotion Detection : A Survey,” IEEE Access, vol. 12, no. December 2023, pp. 18416–18450, 2024, doi: 10.1109/ACCESS.2024.3356357.

[2] Y. Li, J. Chan, G. Peko, and D. Sundaram, “An explanation framework and method for AI-based text emotion analysis and visualisation,” Decis. Support Syst., vol. 178, no. April 2023, p. 114121, 2024, doi: 10.1016/j.dss.2023.114121.

[3] N. Merayo, J. Vegas, C. Llamas, and P. Fernández, “Social Network Sentiment Analysis Using Hybrid Deep Learning Models,” Appl. Sci., vol. 13, no. 20, 2023, doi: 10.3390/app132011608.

[4] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud, “Text-Based Emotion Recognition Using Deep Learning Approach,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/2645381.

[5] I. Shahin, A. B. Nassif, and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network,” IEEE Access, vol. 7, pp. 26777–26787, 2019, doi: 10.1109/ACCESS.2019.2901352.

[6] S. Yang, “Emotion Detection and Analysis Techniques Based on NLP,” in Proceedings of the 3rd International Conference on Mechatronics and Smart Systems, pp. 147–153, 2025, doi: 10.54254/2755-2721/135/2025.21215.

[7] H. Ahmad and M. A. Habib, “Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages,” IEEE Access, vol. 13, no. January, pp. 7538–7554, 2025, doi: 10.1109/ACCESS.2025.3527710.

[8] M. J. Althobaiti, "Arabic Emotion Recognition in Low-Resource Settings: A Novel Diverse Model Stacking Ensemble with Self-Training," Applied Sciences, vol. 13, no. 23, Article no. 12772, 2023, doi: 10.3390/app132312772.

[9] B. Ghasemkhani, K. F. Balbal, and D. Birant, "A New Predictive Method for Classification Tasks in Machine Learning: Multi-Class Multi-Label Logistic Model Tree (MMLMT)," Mathematics, vol. 12, no. 18, Article no. 2825, 2024, doi: 10.3390/math12182825.

[10] U. Malik, S. Bernard, A. Pauchet, C. Chatelain, R. Picot-Clémente, and J. Cortinovis, "Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets," IEEE Access, vol. 12, pp. 15902–15916, 2024, doi: 10.1109/ACCESS.2024.3354705.

[12] M. Irtaza, A. Ali, M. Gulzar, and A. Wali, "Multi-Label Classification of Lung Diseases Using Deep Learning," IEEE Access, vol. 12, pp. 124062–124080, 2024, doi: 10.1109/ACCESS.2024.3454537.

[13] A. Zlatkova, “A Novel CNN-Based Framework for Detection and Classification of Power Quality Disturbances : Exploring Multi-Class Versus Multi- Label Classification,” IEEE Access, vol. 13, pp. 38878–38888, 2025, doi: 10.1109/ACCESS.2025.3546520.

[14] M. Zhang, Y. Yang, S. Qian, Q. Deng, J. Fang, and K. Cai, “ATSIU: A large-scale dataset for spoken instruction understanding in air traffic control,” Adv. Eng. Informatics, vol. 65, pp. 103170, 2025, doi:

10.1016/j.aei.2025.103170.
[15] M. Pilotti, A. M. Waked, A. Mahmoud, H. Hashim, and D. Alzahid, "A dataset of subjective frequency estimates of the words of the Peabody Vocabulary Test by Arabic-English bilingual speakers," Data in Brief, vol. 57, Article no. 111056, Oct. 2024, doi: 10.1016/j.dib.2024.111056.
[16] P. Y. Genest, L. W. Goix, Y. Khalafaoui, E. Egyed-Zsigmond, and N. Grozavu, "French translation of a dialogue dataset and text-based emotion detection," Data Knowl. Eng., vol. 142, pp. 102099, 2022, doi: 10.1016/j.datak.2022.102099.
[17] M. A. Faheem, K. T. Wassif, H. Bayomi, and S. M. Abdou, "Improving neural machine translation for low resource languages through non - parallel corpora : a case study of Egyptian dialect to modern standard Arabic translation," Sci. Rep., pp. 1–10, 2024, doi: 10.1038/s41598-023-51090-4.
[18] J. Alghamdi, Y. Lin, and S. Luo, "Fake news detection in low-resource languages: A novel hybrid summarization approach," Knowledge-Based Systems, vol. 296, Article no. 111884, 2024, doi: 10.1016/j.knosys.2024.111884.
[19] A. Ghosh, S. Umer, B. C. Dhara, and G. G. M. N. Ali, "A Multimodal Pain Sentiment Analysis System Using Ensembled Deep Learning Approaches for IoT-Enabled Healthcare Framework," Sensors, vol. 25, no. 4, Article no. 1223, 2025, doi: 10.3390/s25041223.
[20] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Optimizing Airline Review Sentiment Analysis: A Comparative Analysis of LLaMA and BERT Models through Fine-Tuning and Few-Shot Learning," Comput. Mater. Contin., vol. 82, no. 2, pp. 2769–2792, 2025. doi: 10.32604/cmc.2025.059567.
[21] Q. Guo, C. Wang, D. Xiao, and Q. Huang, "A novel multi-label pest image classifier using the modified Swin Transformer and soft binary cross entropy loss," Engineering Applications of Artificial Intelligence, vol. 126, pt. C, Article no. 107060, 2023, doi: 10.1016/j.engappai.2023.107060.
[22] S. Su, D. Shao, L. Ma, S. Yi, and Z. Yang, "ADCL: An attention feature enhancement network based on adversarial contrastive learning for short text classification," Advanced Engineering Informatics, vol. 65, Article no. 103202, May 2025, doi: 10.1016/j.aei.2025.103202.
[23] B. Zhu and W. Pan, "Chinese text classification method based on sentence information enhancement and feature fusion," Heliyon, vol. 10, no. 17, p. e36861, 2024, doi: 10.1016/j.heliyon.2024.e36861.
[24] J. P. Usuga-Cadavid, B. Grabot, S. Lamouri, and A. Fortin, "Exploring the influence of focal loss on transformer models for imbalanced maintenance data in industry 4.0," IFAC-PapersOnLine, vol. 54, no. 1, pp. 1023–1028, 2021, doi: 10.1016/j.ifacol.2021.08.121.
[25] K. Feng, L. Huang, K. Wang, W. Wei, and R. Zhang, "Prompt-based learning framework for zero-shot cross-lingual text classification," Engineering Applications of Artificial Intelligence, vol. 133, pt. E, Article no. 108481, 2024, doi: 10.1016/j.engappai.2024.108481.
[26] H. Kang, Y. Xu, G. Jin, J. Wang, and B. Miao, "FCAN: Speech emotion recognition network based on focused contrastive learning," Biomedical Signal Processing and Control, vol. 96, pt. B, Article no. 106545, 2024, doi: 10.1016/j.bspc.2024.106545.
[27] J. J. Levett, A. Alnasser, L. M. Elkaim, J. Drager, and T. Pauyo, "Negative sentiments toward anterior cruciate ligament injury prevention outweigh positive awareness discussions on social media," J. ISAKOS, vol. 9, no. 5, p. 100306, 2024, doi: 10.1016/j.jisako.2024.100306.
[28] N. S. Mullah, W. M. N. Wan Zainon, and M. N. Ab Wahab, "Transfer learning approach for identifying negative sentiment in tweets directed to football players," Engineering Applications of Artificial Intelligence, vol. 133, pt. D, Article no. 108377, 2024, doi: 10.1016/j.engappai.2024.108377.
[29] M. A. Soomro, R. N. Memon, A. A. Chandio, M. Leghari, and M. H. Soomro, "A dataset of Roman Urdu text with spelling variations for sentence level sentiment analysis," Data in Brief, vol. 57, Article no. 111170, 2024, doi: 10.1016/j.dib.2024.111170.
[30] P. Ekman, "An Argument for Basic Emotions," Cogn. Emot., vol. 6, no. 3–4, pp. 169–200, 1992, doi: 10.1080/02699939208411068.
[31] F. B. Oliveira, D. Mougouei, A. Haque, J. S. Sichman, H. K. Dam, S. Evans, A. Ghose, and M. P. Singh, "Beyond fear and anger: A global analysis of emotional response to Covid-19 news on Twitter," Online Social Networks and Media, vol. 36, Article no. 100253, 2023, doi: 10.1016/j.osnem.2023.100253.
[32] K. Ullah, I. Mumtaz, M. A. Zia, and A. Razzaq, "Text Based Emotion Detection by Using Classification and Regression Model," in Proc. 16th Int. Conf. Management Science and Engineering Management (ICMSEM 2022), In: J. Xu, F. Altiparmak, M. H. A. Hassan, F. P. García Márquez, and A. Hajiyev, Eds., Lecture Notes on Data Engineering and Communications Technologies, vol. 144. Cham, Switzerland: Springer, 2022, pp. 414–419, doi: 10.1007/978-3-031-10388-9_30.
[33] K. R. Scherer, "What are emotions? and how can they be measured?," Soc. Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005, doi: 10.1177/0539018405058216.
[34] F. Ofli, F. Alam, and M. Imran, "Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response," in Proc. 17th Int. Conf. Information Systems for Crisis Response and Management (ISCRAM), 2020.

[35] R. S. Lazarus, “Thoughts on the Relations Between Emotion and Cognition,” Am. Psychol., vol. 37, no. 9, pp. 1019–1024, 1982.
[36] R. Plutchik, “Emotions and life: Perspectives from psychology, biology, and evolution.,” Am. Psychol. Assoc., 2003.
[37] L. F. Barrett, "Are Emotions Natural Kinds?" Perspectives on Psychological Science, vol. 1, no. 1, pp. 28–58, 2006, doi: 10.1111/j.1745-6916.2006.00003.x.
[38] J. Panksepp, “Affective consciousness: Core emotional feelings in animals and humans,” Conscious. Cogn., vol. 14, no. 1, pp. 30–80, 2005, doi: 10.1016/j.concog.2004.10.004.
[39] A. Maazallahi, M. Asadpour, and P. Bazmi, “Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models,” Inf. Process. Manag., vol. 62, no. 2, pp. 103974, 2025, doi: 10.1016/j.ipm.2024.103974.
[40] A. Jia, Y. Zhang, S. Uprety, and D. Song, “Learning interactions across sentiment and emotion with graph attention network and position encodings,” Pattern Recognit. Lett., vol. 180, pp. 33–40, 2024, doi: 10.1016/j.patrec.2024.02.013.
[41] 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, pt. A, Article no. 118534, 2023, doi: 10.1016/j.eswa.2022.118534.
[42] Y. Ren, Y. Zheng, W. Zhang, D. Qing, X. Zeng, and G. Li, "A novel ensemble over-sampling approach based Chebyshev inequality for imbalanced multi-label data," Neurocomputing, vol. 612, Article no. 128717, 2025, doi: 10.1016/j.neucom.2024.128717.
[43] W. Ren, Y. Zheng, W. Zhang, D. Qing, X. Zeng, and G. Li, “A novel ensemble over-sampling approach based Chebyshev inequality for imbalanced multi-label data,” Neurocomputing, vol. 612, no. August 2023, pp. 128717, 2025, doi: 10.1016/j.neucom.2024.128717.
[44] B. Peng, K. Han, L. Zhong, S. Wu, and T. Zhang, “A head-to-head attention with prompt text augmentation for text classification,” Neurocomputing, vol. 595, 127815, 2024, doi: 10.1016/j.neucom.2024.127815.
[45] S. Radhika, A. Prasanth, and K. K. D. Sowndarya, "A Reliable speech emotion recognition framework for multi-regional languages using optimized light gradient boosting machine classifier," Biomedical Signal Processing and Control, vol. 105, Article no. 107636, 2025, doi: 10.1016/j.bspc.2025.107636.
[46] A. J. Lak, R. Boostani, F. A. Alenizi, A. S. Mohammed, and S. M. Fakhrahmad, "RoBERTa, ResNeXt and BiLSTM with self-attention: The ultimate trio for customer sentiment analysis," Applied Soft Computing, vol. 164, Article no. 112018, 2024, doi: 10.1016/j.asoc.2024.112018.
[47] S. Buechel and U. Hahn, "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis," in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL): Volume 2, Short Papers, Valencia, Spain, Apr. 2017, pp. 578–585, doi: 10.18653/v1/E17-2092.
[48] Y. Li, M. Du, R. Song, X. Wang, M. Sun, and Y. Wang, "Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation," Artificial Intelligence, vol. 332, Article no. 104143, 2024, doi: 10.1016/j.artint.2024.104143.
[49] M. E. Moussa, E. H. Mohamed, and M. H. Haggag, “A generic lexicon-based framework for sentiment analysis,” Int. J. Comput. Appl., vol. 42, no. 5, pp. 463–473, 2020, doi: 10.1080/1206212X.2018.1483813.
[50] Z. Li, L. Chen, Y. Jian, H. Wang, Y. Zhao, M. Zhang, K. Xiao, Y. Zhang, H. Deng, and X. Hou, "Aggregation or separation? Adaptive embedding message passing for knowledge graph completion," Information Sciences, vol. 691, Article no. 121639, 2025, doi: 10.1016/j.ins.2024.121639.
[51] D. S. Chauhan, G. V. Singh, A. Arora, A. Ekbal, and P. Bhattacharyya, "An emoji-aware multitask framework for multimodal sarcasm detection," Knowledge-Based Systems, vol. 257, Article no. 109924, 2022, doi: 10.1016/j.knosys.2022.109924.
[52] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning," in Proc. 29th Int. Conf. Computational Linguistics (COLING), Gyeongju, Republic of Korea, 2022, pp. 4336–4349.
[53] H. Adamu, S. L. Lutfi, N. H. A. H. Malim, R. Hassan, A. Di Vaio, and A. S. A. Mohamed, “Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning,” Sustain., vol. 13, no. 6, 2021, doi: 10.3390/su13063497.
[54] W. Nekoto et al., "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages," in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, pp. 2144–2160, doi: 10.18653/v1/2020.findings-emnlp.195.
[55] A. E. Fagoroye, “University of Ibadan Sentiment Analysis of Low-Resource Yorùbá Tweets Using Fine-Tuned Bert Models,” Journal of Science and Logics in ICT Research, vol. 12, no. 1, pp. 110–122, 2024.
[56] N. Abou Baker, N. Zengeler, and U. Handmann, “A Transfer Learning Evaluation of Deep Neural Networks for Image Classification,” Mach. Learn. Knowl. Extr., vol. 4, no. 1, pp. 22–41, 2022, doi: 10.3390/make4010002.
[57] Y. Kit and M. M. Mokji, “Sentiment Analysis using Pre-trained Language Model with no Fine-tuning and

Less Resource," IEEE Access, vol. 10, pp. 107056–107065, 2022, doi: 10.1109/ACCESS.2022.3212367.
[58] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo, "Cross-lingual pre-training based transfer for zero-shot neural machine translation," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 115–122, 2020, doi: 10.1609/aaai.v34i01.5341.
[59] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, 2019.
[60] A. V Kotelnikova, S. V. Vychegzhanin, and E. V. Kotelnikov, "Cross-Domain Sentiment Analysis Based on Small In-Domain Fine-Tuning," IEEE Access, vol. 11, pp. 41061–41074, 2023, doi: 10.1109/ACCESS.2023.3269720.
[61] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," IEEE Access, vol. 11, pp. 15996–16020, 2023, doi: 10.1109/ACCESS.2022.3224136.
[62] 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, pt. A, Article no. 118534, 2023, doi: 10.1016/j.eswa.2022.118534.
[63] R. Catelli, L. Bevilacqua, N. Mariniello, V. Scotto di Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito, "Cross lingual transfer learning for sentiment analysis of Italian TripAdvisor reviews," Expert Systems with Applications, vol. 209, Article no. 118246, 2022, doi: 10.1016/j.eswa.2022.118246.
[64] N. S. Mullah, W. M. N. Wan Zainon, and M. N. Ab Wahab, "Transfer learning approach for identifying negative sentiment in tweets directed to football players," Engineering Applications of Artificial Intelligence, vol. 133, pt. D, Article no. 108377, 2024, doi: 10.1016/j.engappai.2024.108377.
[65] X. Zheng, Z. Yu, L. Chen, F. Zhu, and S. Wang, "Multi-label Contrastive Focal Loss for Pedestrian Attribute Recognition," in Proc. 25th Int. Conf. Pattern Recognition (ICPR), 2021, pp. 7349–7356, doi: 10.1109/ICPR48806.2021.9413265.
[66] N. Ghanooni, B. Pajoum, H. Rawal, S. Fellenz, V. N. L. Duy, and M. Kloft, "Multi-level Supervised Contrastive Learning," arXiv preprint arXiv:2502.02202, 2025, doi: 10.48550/arXiv.2502.02202.
[67] A. J. Almahdi, A. Mohades, M. Akbari, and S. Heidary, "Enhancing cross-lingual hate speech detection through contrastive and adversarial learning," Engineering Applications of Artificial Intelligence, vol. 147, Article no. 110296, 2025, doi: 10.1016/j.engappai.2025.110296.
[68] Y. Liang, Y. Long, Y. Li, J. Liang, and Y. Wang, "Joint framework with deep feature distillation and adaptive focal loss for weakly supervised audio tagging and acoustic event detection," Digit. Signal Process. A Rev. J., vol. 123, p. 103446, 2022, doi: 10.1016/j.dsp.2022.103446.
[69] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 18661–18673.
[70] S. Liang, L. Shou, J. Pei, M. Gong, W. Zuo, X. Zuo, and D. Jiang, "Label-aware Multi-level Contrastive Learning for Cross-lingual Spoken Language Understanding," in Proc. 2022 Conf. Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 9903–9918, doi: 10.18653/v1/2022.emnlp-main.673.
[71] S. Huang, X. Wu, X. Wu, and K. Wang, "Sentiment analysis algorithm using contrastive learning and adversarial training for POI recommendation," Soc. Netw. Anal. Min., vol. 13, no. 1, pp. 1–14, 2023, doi: 10.1007/s13278-023-01076-x.
[72] V. Suresh and D. C. Ong, "Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP), Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 4381–4394, doi: 10.18653/v1/2021.emnlp-main.359.
[73] Y. Ma, H. Huang, D. Luo, S. Zhang, W. Kang and D. Xie, "Focal Contrastive Learning for Palm Vein Authentication," in IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-15, 2023, Art no. 5025315, doi: 10.1109/TIM.2023.3304689.
[74] J. R. Terven, D. M. Cordova-Esparza, A. Ramírez-Pedraza, E. A. Chávez-Urbiola, and J. A. Romero-González, "A comprehensive survey of loss functions and metrics in deep learning," Artificial Intelligence Review, vol. 58, no. 7, Article no. 195, 2025, doi: 10.1007/s10462-025-11198-7
[75] H. Kang, Y. Xu, G. Jin, J. Wang, and B. Miao, "FCAN: Speech emotion recognition network based on focused contrastive learning," Biomedical Signal Processing and Control, vol. 96, pt. B, Article no. 106545, 2024, doi: 10.1016/j.bspc.2024.106545.
[76] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric Loss For Multi-Label Classification," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 82–91, doi: 10.1109/ICCV48922.2021.00015.
[77] T. Wu and S. Yang, "Contrastive Enhanced Learning for Multi-Label Text Classification," Applied

Sciences, vol. 14, no. 19, Article no. 8650, 2024, doi: 10.3390/app14198650.
[78] X. Chen, W. Zhang, S. Pan, and J. Chen, "Solving Data Imbalance in Text Classification With Constructing Contrastive Samples," IEEE Access, vol. 11, pp. 90554–90562, 2023, doi: 10.1109/ACCESS.2023.3306805.
[79] H. -D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning," IEEE Access, vol. 11, pp. 14742–14751, 2023, doi: 10.1109/ACCESS.2023.3244390.
[80] H. Ding, S. Huang, W. Jin, Y. Shan, and H. Yu, "A Novel Cascade Model for End-to-End Aspect-Based Social Comment Sentiment Analysis," Electron., vol. 11, no. 12, pp. 1–19, 2022, doi: 10.3390/electronics11121810.
[81] S. Maldonado, C. Vairetti, K. Jara, M. Carrasco, and J. López, "OWAdapt: An adaptive loss function for deep learning using OWA operators," Knowledge-Based Systems, vol. 280, Article no. 111022, 2023, doi: 10.1016/j.knosys.2023.111022.
[82] S. F. Yilmaz, E. B. Kaynak, A. Koc, H. Dibeklioglu, and S. S. Kozat, "Multi-Label Sentiment Analysis on 100 Languages With Dynamic Weighting for Label Imbalance," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 331–343, Jan. 2023, doi: 10.1109/TNNLS.2021.3094304.
[83] Y. Wang, L. Nie, and S. Member, "Muti-Modal Emotion Recognition via Hierarchical Knowledge Distillation," IEEE Trans. Multimed., vol. 26, pp. 9036–9046, 2024, doi: 10.1109/TMM.2024.3385180.
[84] F. H. Labib, M. Elagamy, and S. N. Saleh, "EmoBERTa-X: Advanced Emotion Classifier with Multi-Head Attention and DES for Multilabel Emotion Classification," Big Data and Cognitive Computing, vol. 9, no. 2, Article no. 48, 2025, doi: 10.3390/bdcc9020048.
[85] A. Audibert, A. Gauffre, and M.-R. Amini, "Multi-Label Contrastive Learning: A Comprehensive Study," arXiv, arXiv:2412.00101, Jan. 2025, doi: 10.48550/arXiv.2412.00101.
[86] M. Hu, D. Xu, K. He, K. Zhao, and H. Zhang, "Cross-subject emotion recognition with contrastive learning based on EEG signal correlations," Biomedical Signal Processing and Control, vol. 104, Article no. 107511, 2025, doi: 10.1016/j.bspc.2025.107511.
[87] R. S. R. Singh and R. K. Sanodiya, "Zero-Shot Transfer Learning Framework for Plant Leaf Disease Classification," IEEE Access, vol. 11, pp. 143861–143880, 2023, doi: 10.1109/ACCESS.2023.3343759.
[88] S. Chen, Y. Pei, Z. Ke, and W. Silamu, "Low-resource named entity recognition via the pre-training model," Symmetry (Basel)., vol. 13, no. 5, pp. 1–15, 2021, doi: 10.3390/sym13050786.
[89] Y. Xu, M. A. Aslam, W. Jun, N. Ahmed, M. I. Zaman, M. Hamza, and S. Aslam, "Improving Arabic Multi-Label Emotion Classification Using Stacked Embeddings and Hybrid Loss Function," IEEE Access, vol. 13, pp. 188939–188954, 2025, doi: 10.1109/ACCESS.2025.3580478.
[90] E. Ben Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric Loss for Multi-Label Classification," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 82–91, 2021, doi: 10.1109/ICCV48922.2021.00015.
[91] Z. Huang, Y. Ma, R. Wang, W. Li, and Y. Dai, "A Model for EEG-Based Emotion Recognition: CNN-Bi-LSTM with Attention Mechanism," Electron., vol. 12, no. 14, 2023, doi: 10.3390/electronics12143188.
[92] L. Chen, J. Song, X. Zhang, and M. Shang, "MCFL: Multi-label contrastive focal loss for deep imbalanced pedestrian attribute recognition," Neural Computing and Applications, vol. 34, no. 19, pp. 16701–16715, Oct. 2022, doi: 10.1007/s00521-022-07300-7
[93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[94] L. Zhuang, W. Lin, Y. Shi, and J. Zhao, "A Robustly Optimized BERT Pre-training Approach with Post-training," in Proc. 20th Chinese National Conference on Computational Linguistics (CCL), Huhhot, China, Aug. 2021, pp. 1218–1227.
[95] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," in Proc. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS Workshops, 2019.
[96] S. Al-Ghamdi, H. Al-Khalifa, and A. Al-Salman, "Fine-Tuning BERT-Based Pre-Trained Models for Arabic Dependency Parsing," Appl. Sci., vol. 13, no. 7, 2023, doi: 10.3390/app13074225.
[97] C. Mao, Z. Man, Z. Yu, X. Wu, and H. Liang, "Burmese Sentiment Analysis Based on Transfer Learning," J. Inf. Process. Syst., vol. 18, no. 4, pp. 535–548, 2022, doi: 10.3745/JIPS.04.0249.
[98] M. A. H. Wadud, M. F. Mridha, J. Shin, K. Nur, and A. K. Saha, "Deep-BERT: Transfer Learning for Classifying Multilingual Offensive Texts on Social Media," Comput. Syst. Sci. Eng., vol. 44, no. 2, pp. 1775–1791, 2023, doi: 10.32604/csse.2023.027841.
[99] A. Audibert, A. Gauffre, and M.-R. Amini, "Exploring Contrastive Learning for Long-Tailed Multi-label Text Classification," in Machine Learning and Knowledge Discovery in Databases: Research Track, European Conference, ECML PKDD 2024, Vilnius, Lithuania, Sep. 2024, pp. 245–261, doi: 10.1007/978-3-031-70368-3_15.

[100] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 Task 1: Affect in Tweets," in Proc. 12th International Workshop on Semantic Evaluation (SemEval-2018), New Orleans, LA, USA, Jun. 2018, pp. 1–17, doi: 10.18653/v1/S18-1001.
[101] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in Proc. 3rd Workshop Abusive Language Online, Florence, Italy, Aug. 2019, pp. 46–57, doi: 10.18653/v1/W19-3506.
[102] E. Ben Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric Loss for Multi-Label Classification," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 82–91, 2021, doi: 10.1109/ICCV48922.2021.00015.