scientific reports

OPEN

A hybrid stacked ensemble learning framework for multilabel text emotion detection

Hassan Adamu[1,4]✉, Masrah Azrifah Azmi Murad[2,4]✉ & Nurul Amelina Nasharuddin[3,4]

Understanding emotions in text is an important part of sentiment analysis, especially in areas like mental health monitoring, customer feedback analysis, hate speech and disaster management. Unlike basic sentiment analysis that classifies text as positive or negative, real human emotions are complex and often overlapping, requiring multi-label classification to accurately capture multiple emotional states within a single input. While transformer-based models have improved performance, challenges persist particularly in low-resource languages and culturally diverse contexts due to the scarcity of annotated data and the difficulty in generalizing in different languages. This study proposes Hyb-Stack, a hybrid stacked ensemble framework that integrates simple stacking and cross-validation stacking, combining predictions from four transformer-base models (BERT, DistilBERT, RoBERTa, and mBERT) using a Random Forest meta-classifier to enhance classification accuracy, adaptability, and cross-lingual generalisation. Hyb-Stack was evaluated on three datasets: a high-resource English corpus (SemEval-2018 Task 1 E-c) and two low-resource corpora, the Bahasa Indonesia hate speech and HaEmoC_V1, a newly constructed Hausa-language emotion corpus developed to address the lack of annotated data for this language. Experimental results demonstrate that mBERT outperforms individual base models, achieving F1-scores of 82.28 (HaEmoC_V1), 85.33 (Bahasa Indonesia), and 88.90 (SemEval-2018 English). Also, the EM-9 ensemble set (BERT + DistilBERT + mBERT) improves performance, yielding F1-scores of 89.48, 88.19, and 90.67 on the respective datasets, which surpasses both individual models and conventional ensemble strategies such as averaging and weighted averaging. These findings highlight the effectiveness of combining multiple transformers with an optimized decision layer to advance multi-label emotion classification on diverse linguistic contexts.

Keywords Multi-label emotion detection, Hybrid stacking, Low-resource NLP, Ensemble learning, Transformer models, Hausa Language

Emotion detection in text has become an essential aspect of sentiment analysis in various domains, including customer feedback analysis, mental health monitoring, and conversational agents[1]. Unlike traditional sentiment analysis, which mainly categorizes texts as positive, negative, or neutral[2,3] emotion classification aims to identify a broader range of human emotions, such as joy, anger, sadness, fear, and surprise[4]. This deeper understanding of emotions is especially useful in areas like disaster management response or therapy support, where emotional context can guide timely and appropriate interventions[5]. By recognizing specific emotions, systems can prioritize urgent messages, offer personalized support, or trigger alerts for human intervention when necessary[6]. This is very important in serious situations where quick and accurate emotional understanding can help save time or even lives. This level of detail enables systems to more accurately interpret user intent, contextual meaning, and emotional states, thereby enhancing their responsiveness and intelligence[7].

With the advancement of Natural Language Processing (NLP), various models have been developed to detect emotions in text, particularly in English and other high-resource languages[8]. Pre-trained transformer models like BERT and RoBERTa have achieved state-of-the-art results in various classification tasks, including sentiment and emotion detection[9]. However, despite these advancements, the benefits of such models are not equally distributed across all languages. Many low-resource languages lack sufficient annotated corpora and NLP tools,

[1]Department of Computer Science, Universiti Putra Malaysia, Jalan Universiti 1, 43400 Serdang, Selangor, Malaysia. [2]Department of Software Engineering and Information System, Universiti Putra Malaysia, Jalan Universiti 1, 43400 Serdang, Selangor, Malaysia. [3]Department of Multimedia, Universiti Putra Malaysia, Jalan Universiti 1, 43400 Serdang, Selangor, Malaysia. [4]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Jalan Universiti 1, 43400 Serdang, Selangor, Malaysia. ✉email: gs66245@student.upm.edu.my; masrah@upm.edu.my

limiting the effectiveness and adaptability of emotion detection systems in multilingual and culturally diverse contexts[10]. This technology gap makes it difficult to create emotion detection tools in areas where they are most needed, like during emergencies or in communities with fewer resources. This has led to the marginalization of speakers of under-represented languages in the adoption of emerging artificial intelligence (AI) technologies.

Multi-label emotion detection increases the difficulty of the task, as a single input instance may simultaneously express multiple emotions. This requires models to accurately identify and classify overlapping emotion labels within the same textual context[11]. Traditional single-label classification struggles with this task, as it cannot capture the difficulty and overlap of human emotions[12]. Moreover, ambiguity in language, cultural variations in emotional expression, and the subjective nature of emotion perception further complicate the tasks[13]. These challenges require smarter methods, like deep learning, which can better understand the emotional meaning behind words[14]. Also, the lack of annotated datasets for multi-label emotion tasks, especially in low-resource languages, limits the performance and generalizability of existing systems[15].

In the NLP domain, ensemble learning has been shown to improve both performance consistency and predictive accuracy[16]. Among the most prominent techniques are bagging, boosting, and stacking[17]. Their process and standard workflow are illustrated in Supplementary Fig. S1. Bagging methods, like Random Forests, aggregate predictions from multiple learners trained on bootstrapped datasets[18]. Boosting methods, such as AdaBoost and XGBoost, focus on sequentially correcting errors made by previous learners[19]. Stacking, in contrast, employs a meta-learner to combine the predictions of several base models, learning how to best integrate their outputs[20,21]. By leveraging diverse models, ensemble approaches often achieve higher accuracy and stability compared to single models. They can also increase computational complexity and require careful tuning to maximize performance[22].

Most ensemble strategies used in emotion detection are either limited to simple voting mechanisms or lack the architectural flexibility to adapt across multiple languages and resource levels[23]. Hybrid ensemble methods, which integrate stacking with other ensemble paradigms, remain underexplored despite their potential to harness both diversity and coordination among base learners[24]. Current research lacks ensemble models that are specifically optimized for multi-label emotion detection while also being adaptable to linguistically diverse contexts. Developing such models could significantly advance emotion detection capabilities, especially in complex and multilingual contexts. This would allow for better handling of overlapping emotions and variations in language use across different communities[25].

To address the limitations of existing models in multilingual and low-resource contexts, this paper proposes Hyb-Stack, a hybrid stacked ensemble framework for multi-label emotion detection. The architecture integrates simple stacking and cross-validation stacking, combining predictions from four transformer-base models (BERT, DistilBERT, RoBERTa, and mBERT) using a Random Forest meta-classifier to enhance classification accuracy, adaptability, and cross-lingual generalisation. This design enables effective emotion detection on both high-resource and low-resource languages, demonstrating strong adaptability to diverse linguistic and cultural contexts. Beyond its architectural novelty, Hyb-Stack provides a scalable and modular framework for building emotion recognition systems in both monolingual and multilingual environments. This flexibility allows researchers and practitioners to easily extend or customize the framework according to specific language needs or dataset characteristics. Such adaptability is crucial for addressing the wide variability found in emotional expressions on different languages and cultures.

This paper is structured as follows: "Related work" section provides a comprehensive review of related work in emotion detection and ensemble learning, establishing the foundational context for our research. "Methodology" section elaborates on the proposed Hyb-Stack methodology, including its architecture and key innovations. "Experimental results and analysis" section present the experimental setup, results, and a detailed analysis of the findings. Finally, "Conclusion" section concludes the paper by discussing the limitations of the current study and outlining potential directions for future research.

Related work

Emotion detection has evolved significantly, moving from traditional machine learning approaches to deep learning and transformer-based methods[4]. Early research primarily focused on sentiment polarity (positive, negative, neutral), but emotion classification requires finer distinctions across multiple categories such as joy, anger, sadness, fear, and surprise[26]. Understanding these categories is essential because many real-world applications, including customer support, therapy assistance, and disaster management, demand systems that capture emotional representation rather than broad sentiment categories. Frameworks like Plutchik's Emotion Wheel shown in Supplementary Fig. S2 provide theoretical grounding, illustrating how primary emotions combine to form complex states such as optimism or remorse[27]. These psychological models highlight the importance of designing computational systems that can recognize co-occurring or overlapping emotions, especially in multilingual and noisy contexts.

Traditional machine learning methods such as Support Vector Machines (SVMs), Naïve Bayes, and Logistic Regression were widely applied in early emotion detection tasks[28,29]. While effective for basic sentiment classification, they relied heavily on standard features and lexicons, which limited adaptability across domains and languages[30]. To improve performance, researchers turned to ensemble methods such as bagging, boosting, and stacking[17]. Bagging techniques like Random Forests aggregated multiple learners to reduce overfitting[31], while boosting approaches such as AdaBoost and XGBoost iteratively corrected classification errors[32]. Stacking combined base learners through a meta-classifier, offering improved generalization[20]. Although these methods provided performance gains, they struggled with fine-grained emotion recognition and multi-label dependencies, where multiple emotions may co-occur in a single text[33].

The introduction of deep learning marked a turning point for emotion detection. Convolutional Neural Networks (CNNs) demonstrated strong capabilities in capturing local textual features, while Recurrent Neural

Networks (RNNs) and their variants such as LSTMs and GRUs excelled at modelling sequential dependencies[34]. Hybrid deep ensembles further improved performance by combining multiple architectures. For example, ensembles of CNNs and LSTMs captured different aspects of emotional semantics, improving robustness across diverse datasets[35]. Despite these advances, traditional deep models faced limitations such as high computational costs, difficulties with long-range dependencies, and reduced generalization to unseen linguistic or cultural contexts[36]. Moreover, most studies focused on single-label emotion recognition, which oversimplifies the complex and overlapping nature of human affect[15].

Transformer-based architectures revolutionized natural language processing and pushed emotion detection to new levels of accuracy. Models such as BERT, RoBERTa, and XLNet, leveraging self-attention mechanisms, enabled richer contextual representation learning[37]. These models steadily outperformed traditional deep learning systems on benchmark emotion datasets. Lightweight alternatives like DistilBERT[38] and fine-tuning techniques such as adapters offered efficiency gains, though often at the cost of reduced performance in complex tasks. Ensemble learning with transformer[38] emerged as a strategy to further enhance reliability and reduce variance. Studies showed that stacking multiple fine-tuned transformer variants improved generalization and robustness[39]. However, most transformer-based ensembles to date have been restricted to high-resource languages like English and Chinese, single-label context, and computationally expensive pipelines[25].

Multi-label emotion detection introduces unique challenges that further expose the limitations of existing models. A single sentence can express multiple overlapping emotions, requiring systems to capture label correlations and dependencies[11]. Traditional methods such as Binary Relevance treated labels independently, while classifier chains and label powerset methods modelled interdependencies but suffered from scalability and overfitting in high-dimensional label spaces[40,41]. Ensemble strategies like RAkEL improved performance by decomposing label spaces into random subsets[42], yet their reliance on shallow base learners restricted adaptability. Deep ensembles for multi-label emotion detection, including CNN–LSTM hybrids, demonstrated improved accuracy but were rarely tested in low-resource languages. Data scarcity, cultural variations in emotional expression, and class imbalance continue to limit the effectiveness of existing approaches[24,43,44].

Recent work has explored multilingual transformers such as mBERT and XLM-R to extend emotion detection to low-resource languages[45]. These models exploit cross-lingual transfer learning, allowing systems trained on high-resource languages to generalize to underrepresented ones. While promising, existing multilingual approaches often overlook ensemble diversification, and stacking is rarely explored beyond shallow classifiers[20]. Furthermore, meta-learners in stacking commonly logistic regression struggle to capture non-linear dependencies among labels, limiting performance in multi-label contexts[24].

Table 1 outlines key studies on ensemble and hybrid models, highlighting their base models, datasets, target languages, tasks, performance metrics, and key findings. It emphasizes the diversity of approaches and identifies gaps in multi-label and low-resource contexts addressed in this study.

The reviewed studies in Table 1 demonstrate considerable progress in ensemble and deep learning methods for emotion and sentiment analysis, but important gaps remain. A major limitation is the dominance of English, with only a few works extending to Turkish and Chinese, leaving low-resource languages such as Hausa largely unexplored. Research has also concentrated more on sentiment polarity classification than on fine-grained or multi-label emotion recognition, despite evidence that real-world texts often convey overlapping emotions. This imbalance restricts the adaptability of existing models in contexts where multiple emotions co-occur, such as social media communication.

Authors	Ensemble type	Base models	Dataset	Language	Task	Performance metrics
[46]	Deep Transfer Learning	RNN (customized), Sent2Affect	6 benchmark datasets	English	Text-based Emotion Recognition	F1 score = 70
[47]	Joint-Encoder (not ensemble)	Transformer + Co-attention	CMU-MOSEI	English	Emotion + Sentiment	F1 score = 88.19
[48]	Dual-Modality Fusion	LSTM + Sentiment Vectors	UCI Sentiment, Suicide Notes	English	Emotion Classification	Accuracy = 86.3%
[49]	Stacking	CNN, LSTM, BiLSTM, GRU + LR	Real-world ABSA datasets	English	Aspect-Based Sentiment Analysis	Accuracy = 69%
[50]	Transformer + Adapter Layers	Transformer	CMU-MOSEI	English	Emotion Detection	Accuracy = 95%
[51]	Not ensemble	BERT, RoBERTa, ELECTRA, XLM-R	Harnessing Twitter	English	Fine-Grained Emotion Detection	Accuracy = 93.41%
[52]	Neural Ensemble	BERT, ALBERT, RoBERTa, 2D CNN, Wav2Vec 2.0	MELD	English	Bimodal Sentiment Analysis	Accuracy = 77.0
[53]	Hybrid Deep Model	XLNet, BiGRU, Attention	IEMOCAP, CASIA	English, Chinese	Text Emotion Recognition	F-measure = 89.23 Accuracy = 91.71%
[54]	Ensemble (EmoDNN)	Deep NNs + Dropout CNNs	Short Texts	English	Multi-label Emotion Recognition	F-measure = 85.8 Accuracy = 85.8%
[55]	Transfer Learning	RoBERTa-large	11 emotion datasets	English	Emotion Detection	F1-score = 0.84
[56]	Hybrid Attention Model	BERT, CNN, BiGRU + Att	Hotel Reviews, Weibo	Chinese	Sentiment Analysis	F1-score = 88.1%
[57]	Not ensemble (semi-supervised)	CorMulT Transformer	CMU-MOSEI	English	Multimodal Sentiment Analysis	F1-score = 86.7%

Table 1. Summary of representative ensemble and deep learning approaches for emotion and sentiment analysis.

Authors	Datasets	High-resource language	Low-resource language	Train	Dev	Test	Total samples
[60]	SemEval-2018 Task 1 (Task E-c)	English	–	6838	886	3259	10,983
[61]	Indonesia Abusive and Hate Speech Twitter Text	–	Malaya–Indonesia	–	–	–	3293
Our work	Hausa Emotion Corpus (HaEmo_V1)	–	Hausa	–	–	–	12,811

Table 2. Description of the datasets used in this work.

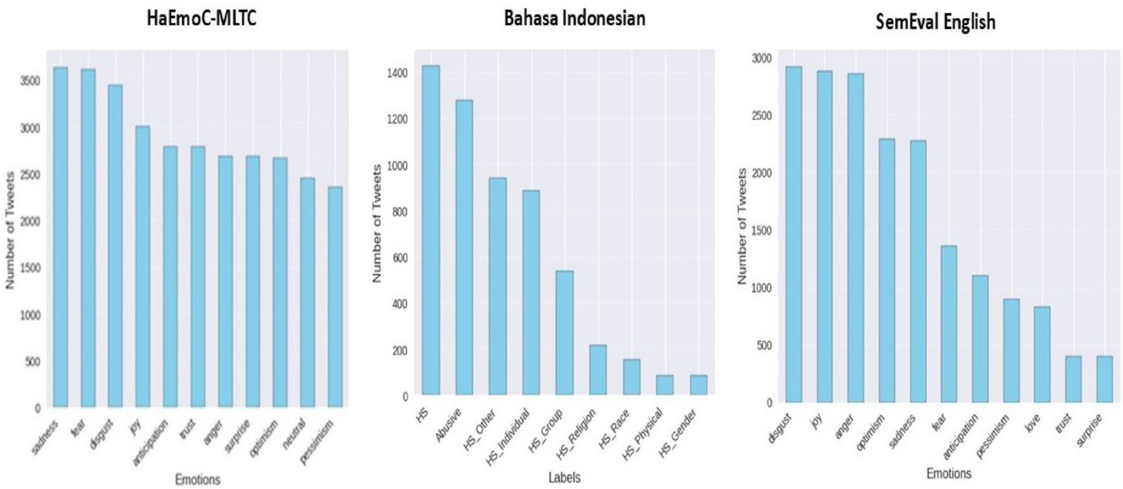


Fig. 1. Distribution of labels in datasets.

To address these challenges, this study proposes Hyb-Stack, a hybrid stacked ensemble framework that combines simple stacking with cross-validation-based stacking to integrate diverse transformer models, including BERT, RoBERTa, DistilBERT, and mBERT. A Random Forest meta-learner is employed to capture non-linear label dependencies within the ensemble. The effectiveness of Hyb-Stack is evaluated on both high-resource (English) and low-resource (Hausa and Bahasa Indonesia) datasets, demonstrating its scalability and adaptability for multilingual emotion detection tasks. The proposed Hyb-Stack ensemble learning architecture is presented in "Methodology" section.

Methodology

In this section, we present a systematic methodology for developing multi-label emotion classification system using transformer-based models and ensemble learning techniques. The system is designed with both high-resource and low-resource text emotion classification datasets strategically partitioned into training, validation, and testing sets. Four transformer architectures were selected as base models based on their complementary characteristics: BERT provides robust bidirectional context understanding through its original transformer architecture[58], while RoBERTa builds upon this foundation with optimized pre-training procedures[59]. The system also incorporates DistilBERT as a computationally efficient alternative that maintains competitive performance through knowledge distillation[38], alongside mBERT which, despite its multilingual design, is included to evaluate architectural consistency in English emotion classification[37]. The methodology follows a structured pipeline consisting of data partitioning, feature embedding, base model training, and comprehensive performance evaluation as shown in Fig. 4.

Dataset description

This study utilizes three datasets encompassing both high-resource and low-resource languages to evaluate the effectiveness of the proposed hybrid stacked ensemble learning framework (Hyb-Stack). The inclusion of diverse linguistic and cultural contexts enables a comprehensive and robust assessment of the model's performance. Table 2 summarizes the datasets, specifying their language categories, sizes, and sources. Notably, the Hausa Emotion Corpus (HaEmoC_V1) was specifically developed in this study to address the paucity of annotated Hausa emotion datasets available for natural language processing (NLP) tasks.

Figure 1 presents the label distributions for the three datasets, HaEmoC-V1 (Hausa), Bahasa Indonesia Hate Speech Dataset, and the SemEval-2018 Task 1 English Emotion Dataset, with the visual comparison highlighting both the internal characteristics of each dataset and the cross-linguistic differences in emotional or toxicity-related expression across platforms.

The three plots illustrated in Fig. 1 reveal that each dataset embodies unique statistical and linguistic characteristics as HaEmoC-V1 provides broad emotional diversity, the Indonesia dataset concentrates on varying intensities and targets of abusive content, and SemEval offers a moderately imbalanced yet rich emotional spectrum, distinctions which underscore the importance of tailoring preprocessing, modelling strategies, and evaluation approaches to the essential properties of each dataset.

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. 2, outlines the key steps involved in data collection, pre-processing, verification, annotation and final organisation.

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. First, text filtering was applied by removing URLs, hashtags, numbers, punctuation marks, and non-Hausa words to focus on the core textual content. Next, keyword-based filtering was used to prioritize tweets containing specific keywords related to disaster relief and emergency response (e.g., Tallafin Gaggawa, Tallafin kayan abinci, Tallafin Annoba), thereby enhancing relevance. Finally, tokenization and stopword removal were carried out, where the text was split into individual words or tokens, and common Hausa stopwords (e.g., da, ta, shi) were removed to eliminate non-informative elements.

C. 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.

D. 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.

E. 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

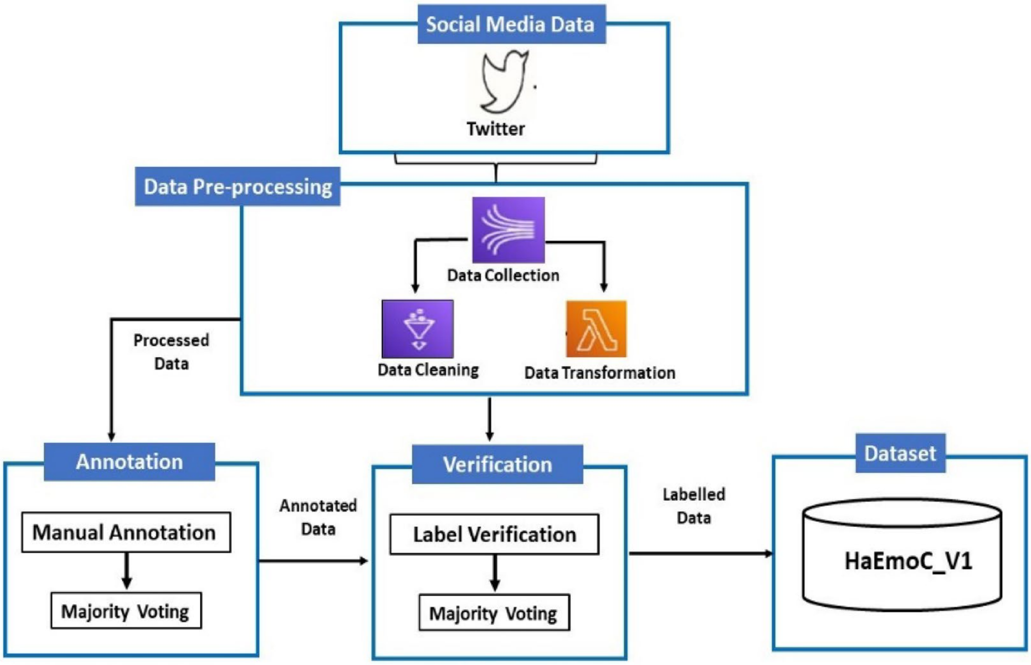


Fig. 2. Dataset acquisition framework.

emotion classification, ensuring a high-quality benchmark for low-resource language sentiment analysis. The entire dataset preparation framework is illustrated in Fig. 2.

HaEmoC_V1 statistics

This section provides a detailed statistical overview of the Hausa Emotion Corpus (HaEmoC_V1), a multi-label dataset designed for emotion classification in the Hausa language. 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 crucial role in benchmarking transformer-based models for low-resource language processing, as shown in Table 3.

Sample of HaEmoC-V1 dataset

Table 4 presents some samples of the instances from the newly constructed (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. These examples help illustrate the dataset's suitability for multi-label emotion classification in low-resource NLP tasks. Figure 3 illustrates the statistical analysis of the emotions in each category.

Dataset correlation

This section reports the correlation analysis conducted on the HaEmoC_V1 dataset to examine the interrelationships among emotion labels. Figure 3 presents the correlation matrix, where darker shades denote stronger associations between emotions. In particular, trust and optimism (0.9) as well as pessimism and anger (0.8) exhibit strong positive correlations, whereas joy and sadness (– 0.32) show a moderate negative correlation. These findings provide valuable insights into patterns of emotion co-occurrence, which are particularly relevant for enhancing the performance of multi-label emotion classification models.

Transformer-based modelling framework

The proposed architecture illustrated in Fig. 4 presents a multi-label emotion classification framework that accommodates both high-resource and low-resource languages. The model utilizes datasets comprising English for high-resource language representation and HaEmoC_V1 and the Indonesia Hate Speech corpus for low-resource languages. The data is initially divided into training, validation, and testing sets. The actual training set undergoes feature embedding and is processed using a Binary Cross-Entropy (BCE) with Logit Loss Function, which is suitable for handling multi-label classification problems. The system then fine-tunes four transformer models specifically BERT, RoBERTa, DistilBERT, and mBERT as base models. Each model is trained independently, and their predictions are stored for ensemble learning approach. The ensemble combines the strengths of individual models to enhance overall performance and generalization. Finally, the performance of the models is evaluated using standard metrics including accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the model's effectiveness across different linguistic contexts.

Baseline transformer models

The baseline transformer models used in this study vary in size and architectural complexity, reflecting differences in their number of parameters and design. BERT-base consists of approximately 110 million parameters, featuring 12 transformer encoder layers, a hidden size of 768, and 12 self-attention heads. DistilBERT, a compressed variant of BERT, contains around 66 million parameters by reducing the number of encoder layers to six while retaining a hidden size of 768 and 12 attention heads. This reduction results in a lighter model, enabling faster training and inference. RoBERTa-base shares a similar architecture with BERT-base but benefits from optimized pre-training strategies, comprising approximately 125 million parameters, 12 layers, a hidden size of 768, and 12 attention heads. mBERT (Multilingual BERT-base) also has roughly 110 million parameters and the same architectural configuration as BERT-base, but it is pre-trained on multilingual corpora covering over 100 languages. These model sizes offer a balance between representational capacity and computational efficiency, facilitating fair comparison and effective ensemble integration across both high-resource and low-resource language scenarios.

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 (1 text column + 11 emotion columns + 1 ID column)

Table 3. Dataset description of Hausa Emotion Corpus (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

Table 4. Sample of HaEmoC-V1. ang, Anger; sad, Sadness; dis, Disgust; fea, Fear; sur, Surprise; joy, Joy; tru, Trust; opt, Optimism; pes, Pessimism; ant, Anticipation; neu, Neutral.

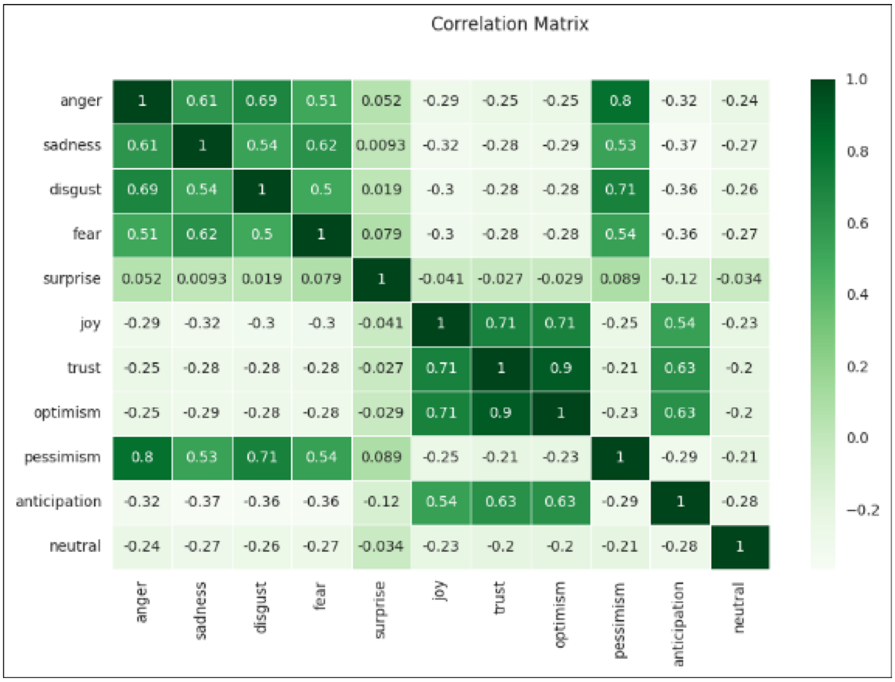


Fig. 3. Dataset correlation.

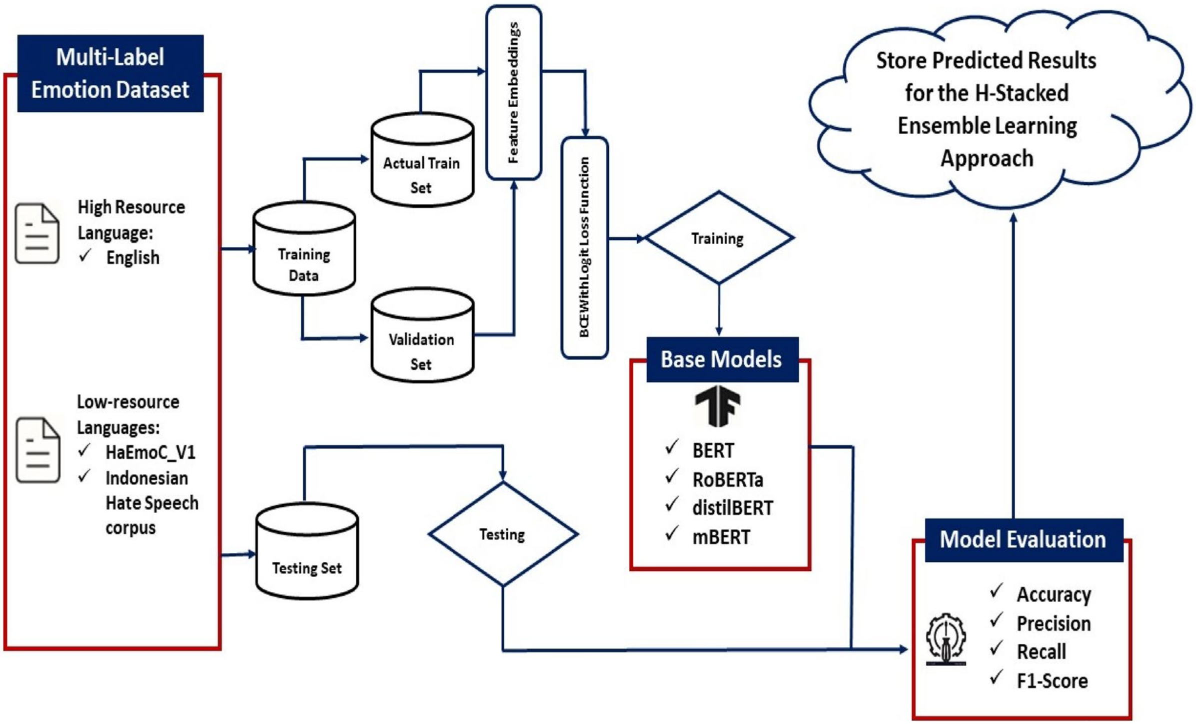


Fig. 4. Modelling of the Transformer-base models.

Ensemble-based approach

After fine-tuning and evaluating the pre-trained transformer models individually as shown in Fig. 4, we applied multiple ensemble strategies including average and weighted average methods alongside our proposed hybrid stack ensemble learning (Hyb-Stack) approach. The Hyb-Stack method combines two stack ensemble techniques: simple stacking and cross-validation-based stacking, enabling a comprehensive comparison of ensemble performance. Hyb-Stack framework consists of three primary layers. First, the base learner layer consists of fine-tuned transformer models that generate probability distributions for each emotion label. Their predictions are then aggregated in the ensemble combination layer, where various strategies were systematically explored to maximize model diversity. Finally, the meta-classifier synthesizes these ensemble outputs to produce the final multi-label predictions. Equation 1 describes how the ensemble sets were generated using a systematic combination of the four models employed in this research.

$$C\left(n,k\right)=\frac{n!}{k!\left(n-k\right)!}, \tag{1}$$

where n is the total number of base models; k is the number of instances.

$$\begin{aligned}&Therefore,\ n=4,\ \text{and}\ k=2,3,4\ which\ yielded:\\&C\left(4,2\right)+C\left(4,3\right)+C\left(4,4\right)=6+4+1=11\end{aligned}$$

Hence, 11 ensemble method (EM) combinations were generated, designated EM-1 to EM-11. Each ensemble configuration generates a set of logits across the predefined emotion classes, and different aggregation techniques are applied for prediction.

Average ensemble approach

In the average ensemble, we average the logits output for each test instance across models. The final predicted class is the one with the maximum average logit value. This is computed as shown in Eq. 2 and Algorithm 1:

$$O=\arg\max_{i\in(1,d)}\left(\sum_{j=1}^{m}\text{bce\_logit}_{ij}\right) \tag{2}$$

where bce_logit_ij represents the raw logit of the jth model for the ith instance. The final class is assigned by taking the index of the maximum averaged value.

```
Inputs:
m[ ] ← models
d[ ] ← test instances
bce_logit[ ] ← logits from BCEWithLogits
sum[ ] ← summed logits
avg[ ] ← average logits
nclass ← number of classes

for i ∈ (1, d) do
    for j ∈ (1, m) do
        sum[i] = sum[i] + bce_logit[i][j];
    end
    avg[i] = sum[i] / m;
end
O = argmax(avg) // Output class indices
```

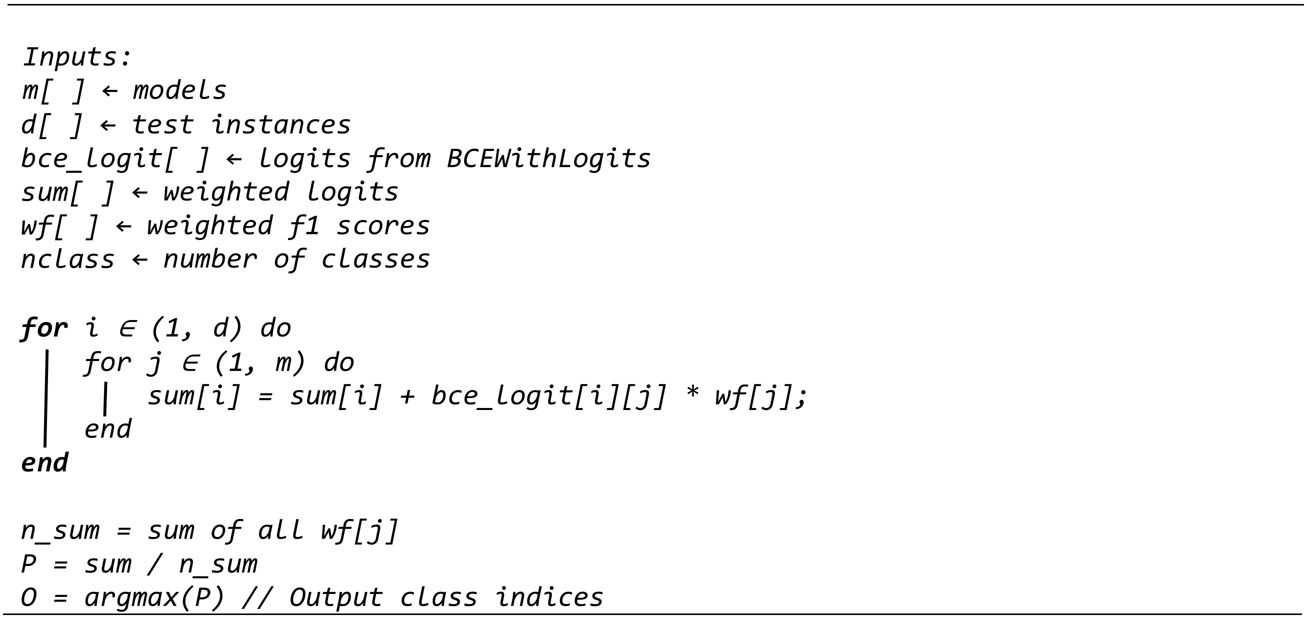
Algorithm 1. Average ensemble method

Weighted average ensemble approach

In the weighted average ensemble, each model's logits are multiplied by its corresponding weighted F1-score before aggregation. The weighted average is presented in Eq. 3 and Algorithm 2:

$$O=\arg\max_{i\in(1,d)}\left(\frac{\sum_{j=1}^{m}\text{bce\_logit}_{ij}\cdot wf_j}{\sum_{j=1}^{m}wf_j}\right) \tag{3}$$

where wf_j is the weighted F1-score of models j, emphasizing high-performing models in the final prediction.



```
Inputs:
m[ ] ← models
d[ ] ← test instances
bce_logit[ ] ← logits from BCEWithLogits
sum[ ] ← weighted logits
wf[ ] ← weighted f1 scores
nclass ← number of classes

for i ∈ (1, d) do
    for j ∈ (1, m) do
        sum[i] = sum[i] + bce_logit[i][j] * wf[j];
    end
end

n_sum = sum of all wf[j]
P = sum / n_sum
O = argmax(P) // Output class indices
```

Algorithm 2. Weighted average ensemble method

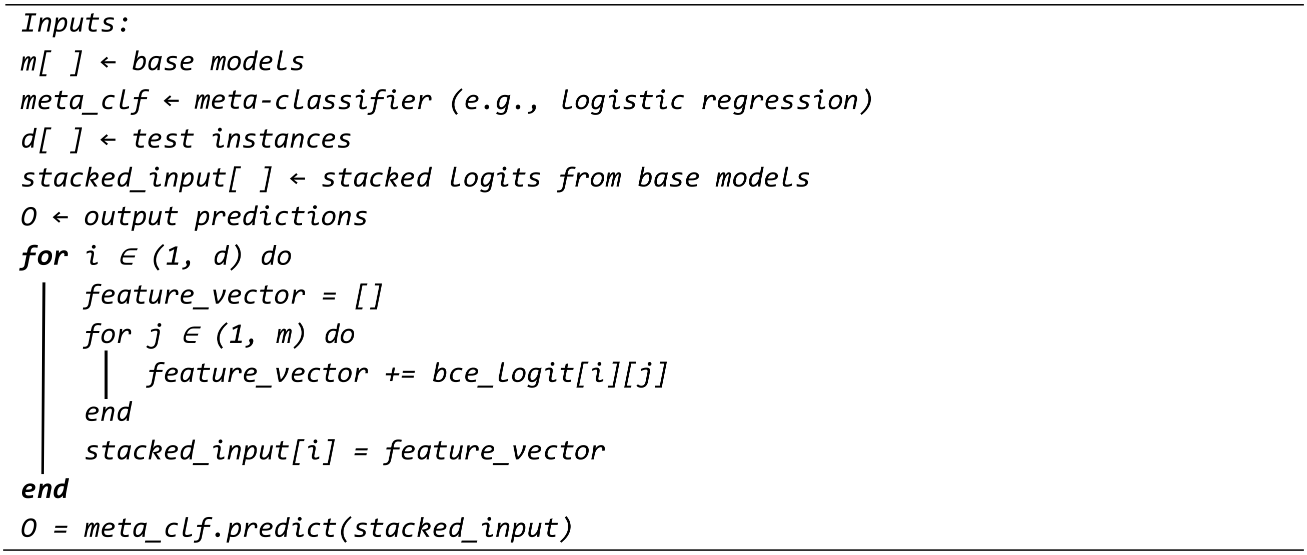
Hybrid stack ensemble (Hyb-Stack) approach

This is the proposed ensemble methods in this research. In this approach, the logits of all base models (BERT, DistilBERT, RoBERTa, and mBERT) are stacked into a feature vector, which is then passed to a meta-classifier specifically Random Forest. This model is trained on validation logits to learn the optimal combination of base model outputs. The Hyb-Stack output is calculated using Eq. 4:

$$O = \text{meta\_clf}\ \left(\text{Concat}\left(\text{bce\_logit}_1, \text{bce\_logit}_2, \ldots, \text{bce\_logit}_m\right)\right) \tag{4}$$

where meta_clf is the trained meta-classifier and Concat represents vector concatenation of logits from all base models.

The full architecture in Fig. 5 and the workflow in Algorithm 3 illustrate the development from base-model predictions to meta-learner improvement, ensuring strong and adaptable emotion classification.



```
Inputs:
m[ ] ← base models
meta_clf ← meta-classifier (e.g., logistic regression)
d[ ] ← test instances
stacked_input[ ] ← stacked logits from base models
O ← output predictions
for i ∈ (1, d) do
    feature_vector = []
    for j ∈ (1, m) do
        feature_vector += bce_logit[i][j]
    end
    stacked_input[i] = feature_vector
end
O = meta_clf.predict(stacked_input)
```

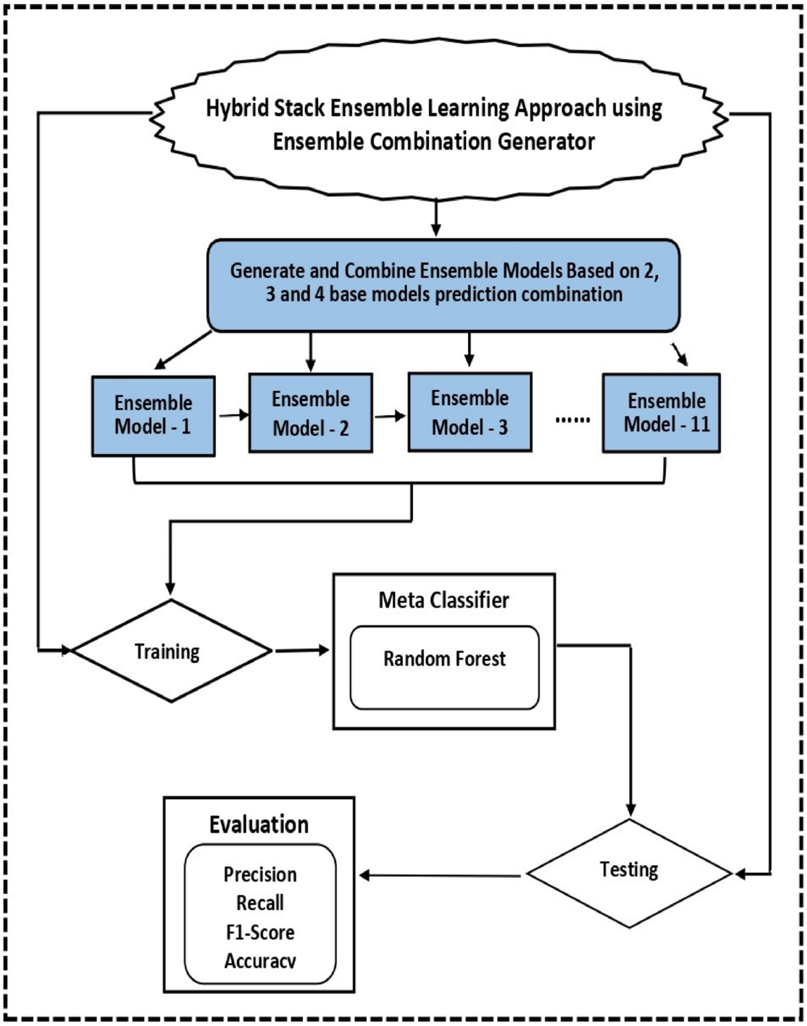

Hybrid Stack Ensemble Learning Approach using Ensemble Combination Generator
Generate and Combine Ensemble Models Based on 2, 3 and 4 base models prediction combination
Ensemble Model - 1
Ensemble Model - 2
Ensemble Model - 3
......
Ensemble Model - 11
Training
Meta Classifier
Random Forest
Testing
Evaluation
Precision
Recall
F1-Score
Accuracv

Fig. 5. Modelling using the proposed ensemble learning approach (Hyb-Stack).

Algorithm 3. Hyb-stack ensemble

Experimental results and analysis

In this section, we present an in-depth analysis of the models' performance on the three datasets used in this study. The aim is to evaluate the effectiveness of the proposed hybrid stacked ensemble learning (Hyb-Stack) approach by comparing its performance with standard transformer models used as base learners, as well as with existing ensemble methods such as average and weighted ensemble learning. The results cover both low-resource and high-resource language datasets, enabling a comprehensive assessment of the model's generalization capabilities. Through detailed error analysis, evaluation of ensemble contributions, and comparisons with prior studies, this section highlights the empirical strengths and practical advantages of the proposed system.

Experimental setup and evaluation metrics

The experiments were conducted on Google Colab Pro using a T4 or V100 GPU based on their availability (16–40 GB VRAM), supported by 16 GB RAM and an Intel Core i7-10110U CPU (3.10 GHz base, 3.59 GHz boost) to accelerate training. The implementation used Python 3.8 with key libraries: PyTorch for deep learning, HuggingFace Transformers for pre-trained language models, and NumPy/scikit-learn for data processing and evaluation. All models were trained with the AdamW optimizer (learning rate = 2e–5) and a batch size of 16, balancing computational efficiency and performance. Training proceeded for 11 epochs with early stopping (monitoring validation loss) to avoid overfitting while ensuring convergence. For text pre-processing, we used each transformer model's native tokenizer, truncating/padding sequences to 128 tokens for uniformity. These hyper-parameters followed established fine-tuning practices while optimizing for available hardware.

For baseline comparisons, we evaluated four widely adopted transformer models: BERT, RoBERTa, DistilBERT, and mBERT, selected for their proven contextual understanding and multilingual performance[33].

Our proposed Hyb-Stack ensemble combined these architectures to leverage their complementary strengths, aiming to enhance generalization and predictive accuracy. Performance was assessed using micro/macro-averaged F1-scores, precision, recall, and accuracy, providing a comprehensive evaluation of both class-wise and overall model behaviour.

Comparative performance of the base models

Table 5 summarizes the performance of individual transformer models on the test datasets, evaluated using weighted F1-score serving as the primary metric for model effectiveness. Table 6 compares ensemble configurations, demonstrating performance improvements achieved through different combination strategies including average, weighted average, and our proposed Hyb-Stack Ensemble.

After completing the experiment of the four transformer-based models on all three datasets as presented in Table 5, the results revealed that mBERT consistently achieves the highest F1-scores, indicating robust cross-lingual performance. On the low-resource HaEmoC_V1 dataset, mBERT attains an F1-score of 82.28%, outperforming BERT (79.65%), DistilBERT (81.84%), and RoBERTa (73.88%). Performance improves for all models on the Bahasa Indonesia Hate Speech dataset, where mBERT again leads with an F1-score of 85.33%, followed closely by DistilBERT (84.93%) and BERT (84.88%), while RoBERTa records a lower score of 79.16%. The highest F1-scores are observed on the SemEval-2018 English dataset, reflecting greater language resource availability, with mBERT achieving 88.90%, followed by DistilBERT (86.44%), RoBERTa (85.66%), and BERT (85.21%). Overall, the progression of F1-scores from Hausa to Bahasa Indonesia and then to English highlights the influence of language coverage in pretrained models and underscores the effectiveness of multilingual representations for emotion and hate speech classification.

Figure 6 illustrates a line chart showing the error rates of each model. The chart highlights clear contrasts in model behaviour. RoBERTa begins with the highest error rate of 26% on HaEmoC_V1 but improves substantially on subsequent datasets, showing adaptability with larger, data-rich resources. mBERT demonstrates the lowest overall error rates, improving steadily to 11% on SemEval-2018, confirming its strong cross-lingual generalization. DistilBERT maintains competitive error rates throughout, consistently outperforming BERT while remaining computationally efficient. BERT shows stable but less dynamic error rates of 15–20% on all the datasets. These trends emphasize that mBERT generalizes best, RoBERTa adapts most dramatically, and DistilBERT balances performance with efficiency.

Training and validation performance

Figure 7 illustrates the training and validation curves for mBERT, the best-performing base model. The curves show a consistent reduction in training and validation loss on all the 11 epochs, accompanied by steadily improving validation accuracy. The absence of divergence between training and validation confirms effective convergence without overfitting, further validating mBERT strong generalization.

Hybrid stack ensemble learning results (Hyb-Stack)

Table 6 summarizes the results of eleven ensemble configurations evaluated with the three datasets using average ensemble, weighted ensemble, and the proposed Hyb-Stack method.

The experimental results in Table 6 demonstrate that the Hyb-Stack ensemble consistently outperforms simple and weighted averaging baselines across all evaluated datasets. By integrating the four transformer models (BERT, RoBERTa, DistilBERT, and mBERT), Hyb-Stack achieves substantial improvements in precision, recall, and F1-score for multilingual and multi-label classification. F1-scores range from 83.05 to 89.48% on the HaEmoC dataset, 81.39% to 88.19% on the BHIS dataset, and 85.13% to 90.67% on the SemEval-2018 English dataset, with the EM-9 configuration (BERT + DistilBERT + mBERT) achieving the highest scores of 89.48%, 88.19%, and 90.67%, respectively. Overall performance improves from Hausa to Bahasa Indonesia and then to English, reflecting differences in language resource availability, while weighted averaging offers moderate gains over simple averaging but remains inferior to the Hyb-Stack approach.

Models	Datasets	Precision	Recall	F1-score	Accuracy
BERT	HaEmoC_V1	80.42	78.39	79.65	84.27
	Bahasa Indonesia Hate Speech	83.56	82.74	84.88	87.19
	Sem-Eval-2018 Task 1 E-c English	85.67	82.48	85.21	89.34
DistilBERT	HaEmoC_V1	82.91	77.36	81.84	87.68
	Bahasa Indonesia Hate Speech	84.77	84.22	84.93	88.34
	Sem-Eval-2018 Task 1 E-c English	87.59	85.15	86.44	94.02
RoBERTa	HaEmoC_V1	75.62	71.93	73.88	81.17
	Bahasa Indonesia Hate Speech	80.84	77.38	79.16	85.92
	Sem-Eval-2018 Task 1 E-c English	86.53	84.79	85.66	91.45
mBERT	HaEmoC_V1	84.37	83.65	82.28	86.73
	Bahasa Indonesia Hate Speech	86.21	83.86	85.33	91.91
	Sem-Eval-2018 Task 1 E-c English	90.13	85.26	88.90	94.68

Table 5. Transformer models result.

			Avg. EM			Wt. Avg. EM			Hyb-Stack ensemble		
EM sets	Models	Datasets	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
EM—1	BERT + RoBERTa	HaEmoC	80.14	78.91	79.62	82.30	81.84	84.51	85.12	83.78	85.27
		BHIS	83.47	82.03	84.69	84.51	84.29	86.90	87.65	85.09	84.96
		SE-18-T1 E-c	85.76	82.65	85.12	86.41	85.98	87.32	88.44	86.73	86.11
EM—2	BERT + DistilBERT	HaEmoC	82.11	80.90	81.55	85.77	83.65	82.12	86.39	85.78	84.41
		BHIS	84.58	83.10	82.94	83.37	82.28	84.51	87.06	84.32	85.78
		SE-18 T1 E-c	87.81	85.12	84.67	86.73	84.65	85.44	88.24	87.95	86.73
EM—3	BERT + mBERT	HaEmoC	83.29	81.41	82.58	85.13	82.66	86.79	87.01	86.88	87.36
		BHIS	84.35	80.13	83.22	84.80	81.11	83.55	86.49	87.63	86.07
		SE-18 T1 E-c	85.78	83.98	84.69	83.03	81.90	84.15	89.74	88.21	87.92
EM—4	RoBERTa + mBERT	HaEmoC	86.52	84.44	83.05	85.90	84.22	83.76	86.88	87.31	85.64
		BHIS	84.26	82.94	81.71	84.00	80.59	84.91	85.82	84.37	83.74
		SE-18 T1 E-c	83.92	81.61	80.35	89.21	87.53	86.17	89.88	88.14	87.33
EM—5	DistilBERT + mBERT	HaEmoC	82.45	79.97	81.02	85.78	84.60	83.12	87.06	82.97	84.59
		BHIS	84.99	82.33	82.88	86.01	83.85	85.42	88.90	84.93	86.15
		SE-18 T1 E-c	87.17	85.29	86.33	88.84	85.90	86.01	89.53	86.90	88.47
EM—6	DistilBERT + RoBERTa	HaEmoC	84.60	80.75	81.64	84.31	83.21	82.59	85.17	84.67	83.92
		BHIS	80.43	78.23	79.99	83.15	80.31	81.58	84.78	85.82	83.17
		SE-18 T1 E-c	87.76	86.20	85.96	87.12	83.38	84.27	86.58	84.71	85.37
EM—7	BERT + RoBERTa + mBERT	HaEmoC	86.98	80.85	83.63	86.74	82.58	84.91	87.24	83.39	85.40
		BHIS	85.05	82.18	83.01	82.72	80.79	81.64	89.13	87.93	86.06
		SE-18 T1 E-c	84.61	81.45	80.22	86.92	84.04	85.13	87.89	85.94	87.67
EM—8	BERT + DistilBERT + RoBERTa	HaEmoC	85.89	82.99	84.70	83.37	82.43	83.11	88.21	88.67	86.23
		BHIS	86.20	84.11	85.94	87.09	86.25	84.44	86.80	85.42	85.58
		SE-18 T1 E-c	83.67	86.71	88.84	87.32	84.07	85.73	89.22	87.33	88.68
EM—9	BERT + DistilBERT + mBERT	HaEmoC	87.79	86.58	85.14	85.61	87.73	86.25	90.91	89.26	89.48
		BHIS	86.23	85.99	84.82	88.67	86.70	86.30	89.43	87.81	88.19
		SE-18 T1 E-c	89.12	87.34	86.41	89.54	87.12	87.73	91.88	89.55	90.67
EM—10	DistilBERT + RoBERTa + mBERT	HaEmoC	83.83	83.40	82.94	84.01	82.36	83.67	87.40	89.23	87.81
		BHIS	83.99	80.86	81.39	85.60	82.54	84.91	86.08	87.02	85.44
		SE-18 T1 E-c	85.06	85.81	86.59	86.77	84.28	86.10	87.02	85.27	86.36
EM—11	BERT + DistilBERT + RoBERTa + mBERT	HaEmoC	82.74	85.80	84.31	83.51	80.33	84.84	86.33	85.65	84.96
		BHIS	84.16	82.40	83.91	88.73	84.83	86.50	88.05	86.27	87.69
		SE-18 T1 E-c	89.53	85.49	86.10	87.93	86.08	85.34	88.86	87.58	89.62

Table 6. Performance of the proposed Hyb-Stack ensemble learning approach. EM, Ensemble Method; Avg. EM, Average Ensemble Method; Wt. Avg. EM, Weighted Average Ensemble Method; Hyb-Stack Ensemble, Hybrid Stacked Ensemble; Prec. , Precision; Rec. , Recall; F1, F1-score, HaEmoC, Hausa Emotion Corpus, BHIS, Bahasa Indonesia Hate Speech Dataset, SE-18-T1 E-c, Sem-Eval-2018 Task 1 E-c English.

Figure 8 presents the error rates of 11 ensemble models (EM-1 to EM-11). The chart shows that the EM-9 ensemble set achieved the highest performance, with the lowest error rate of 10.55%, whereas EM-6 recorded the weakest performance with an error rate of 15.85%. Most models produced moderate error rates, ranging between 12.58% and 14.55%, with EM-11, EM-3, and EM-8 also performing competitively. The substantial performance gap between EM-9 and the other models suggests that its ensemble configuration, including the choice of base models, stacking methodology, and hyperparameter tuning was more effective in minimizing errors.

Confusion matrix

The confusion matrix of the best-performing ensemble model (EM-9) illustrated in Fig. 9 demonstrates strong classification performance, with most predictions concentrated along the diagonal, reflecting high accuracy. Misclassifications were relatively limited and mainly occurred between semantically close emotion classes, such as joy vs. surprise and sadness vs. fear. These confusions highlight the inherent overlap in emotional expressions and illustrate the challenges of multi-label emotion classification.

Comparison with other studies

Recent advances in emotion classification have increasingly relied on ensemble learning to combine the strengths of multiple models. Table 7 compares seven state-of-the-art ensemble approaches, highlighting their architectures, key innovations, and performance metrics. The analysis reveals how strategic model combinations and meta-learning techniques achieve varying success in different languages (e.g., Arabic, Bengali) and domains

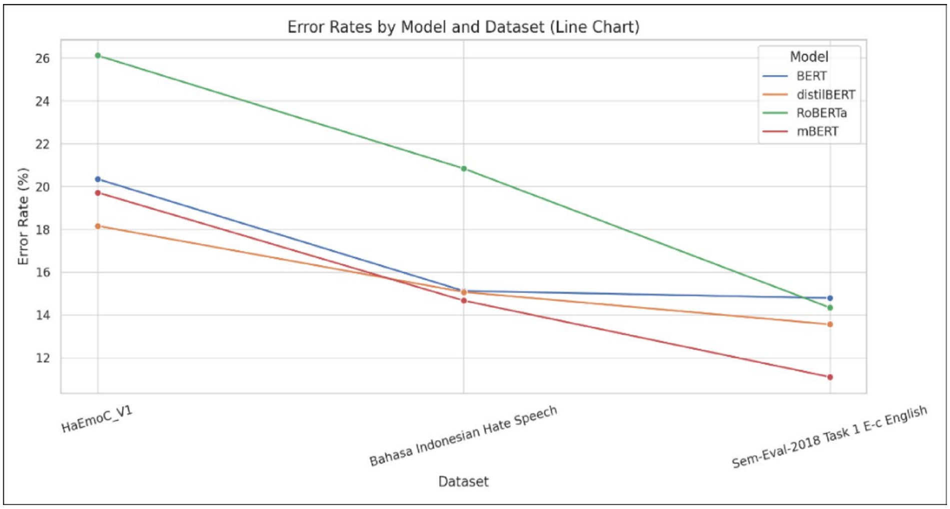


Fig. 6. Error rate of transformer-based models.

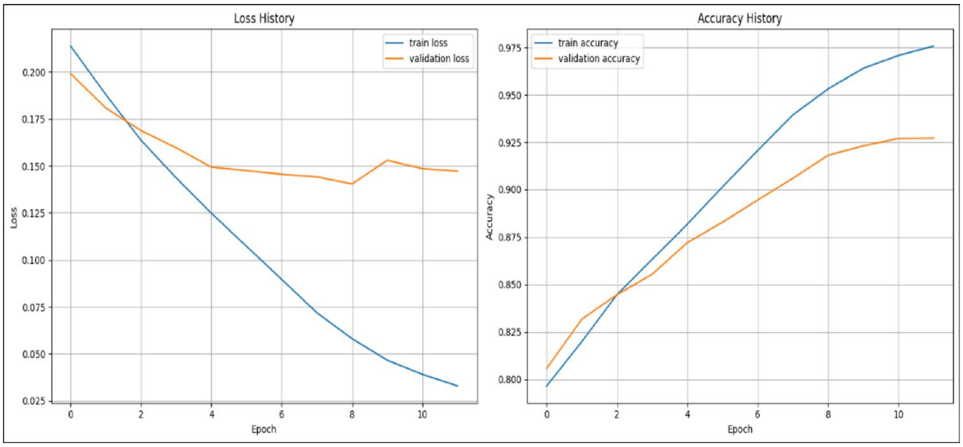


Fig. 7. mBERT model's training and validation performance.

(e.g., social media content such as tweets). Our proposed Hyb-Stack ensemble emerges as the new performance leader, demonstrating the effectiveness of hybrid transformer architectures with advanced ensemble strategies.

Conclusion

This study presented Hyb-Stack, a hybrid stacked ensemble learning framework for multi-label text emotion detection that integrates simple stacking with cross-validation-based stacking. The framework was evaluated on three datasets, HaEmoC_V1 (a newly developed Hausa emotion corpus), Bahasa Indonesia Hate Speech, and SemEval-2018 Task 1 E-c (English) to examine whether ensemble learning can consistently improve performance over individual transformer-based models across languages with differing resource availability. Experimental findings show that the proposed Hyb-Stack framework, particularly the EM-9 ensemble configuration (BERT, DistilBERT, and mBERT), achieved higher macro-F1 scores than the strongest individual baseline across all evaluated datasets. These results indicate that hybrid stacking can enhance predictive performance, with more noticeable benefits in low-resource language settings. The strong performance of mBERT highlights the value of multilingual pre-training for cross-lingual emotion modelling, while the ensemble results suggest that combining complementary transformer representations can yield more robust emotion predictions.

Limitations and future work

While the proposed Hyb-Stack framework demonstrates consistent performance improvements across the evaluated datasets, some scope-related considerations should be noted. The experimental analysis was conducted on three publicly available datasets, which, although diverse in language and domain, do not cover all possible forms of emotional expression. In addition, the study focused on the emotion categories defined within these datasets, and did not explore more fine-grained or context-dependent emotional phenomena. Finally, as with

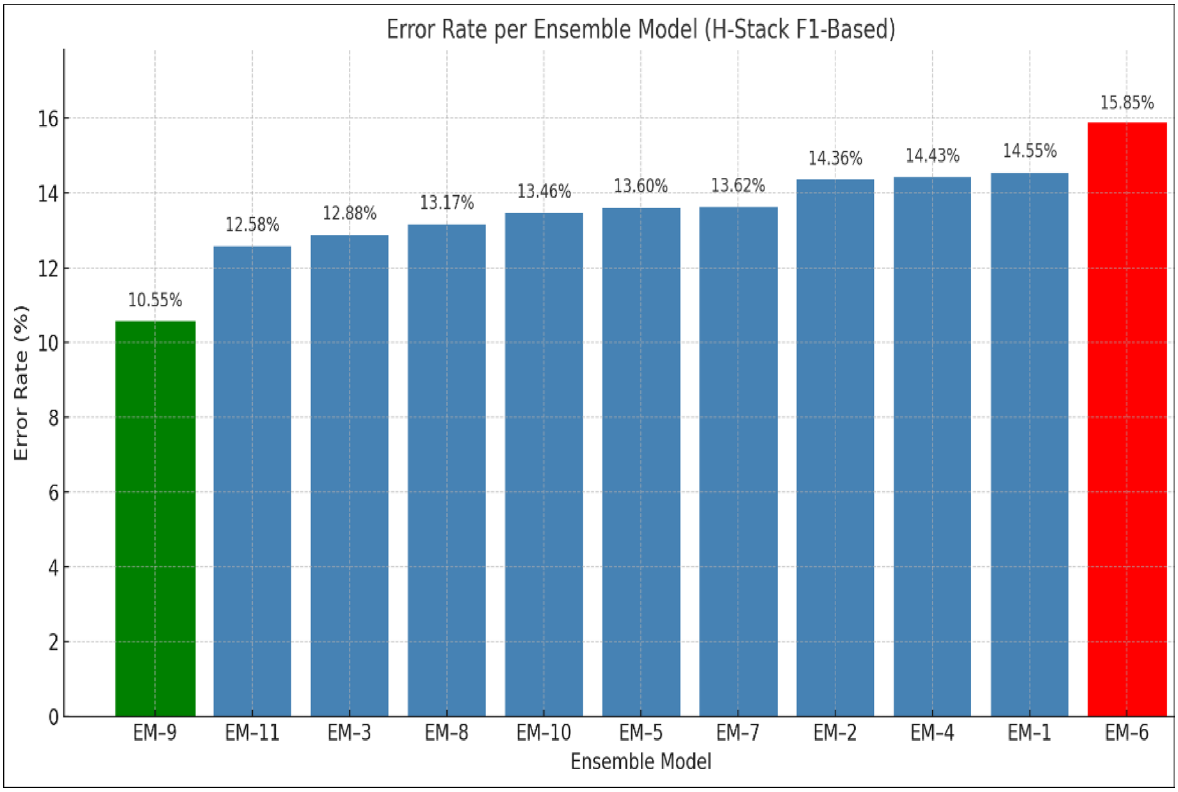


Fig. 8. Error rate of the H-stack ensemble approach.

most ensemble-based approaches, the framework involves additional model components, which may require further optimization for large-scale or real-time applications.

Future research will focus on extending the Hyb-Stack framework to a wider range of emotion categories and further enriching the HaEmoC_V1 corpus with larger and more diverse data sources. Additional studies may explore the framework's ability to process code-mixed and code-switched texts, emoticon- and emoji-based emotional expressions, and mixed-emotion content commonly found in social media. Evaluating the approach on other underrepresented African and Indigenous languages would provide a stronger assessment of its generalizability. Methodological enhancements, such as dynamic ensemble selection, alternative meta-learning strategies, and the integration of larger multilingual transformers (e.g., XLM-R), also represent promising directions. Finally, addressing bias mitigation, interpretability, and deployment considerations will be essential for improving the robustness and real-world applicability of ensemble-based emotion detection systems.

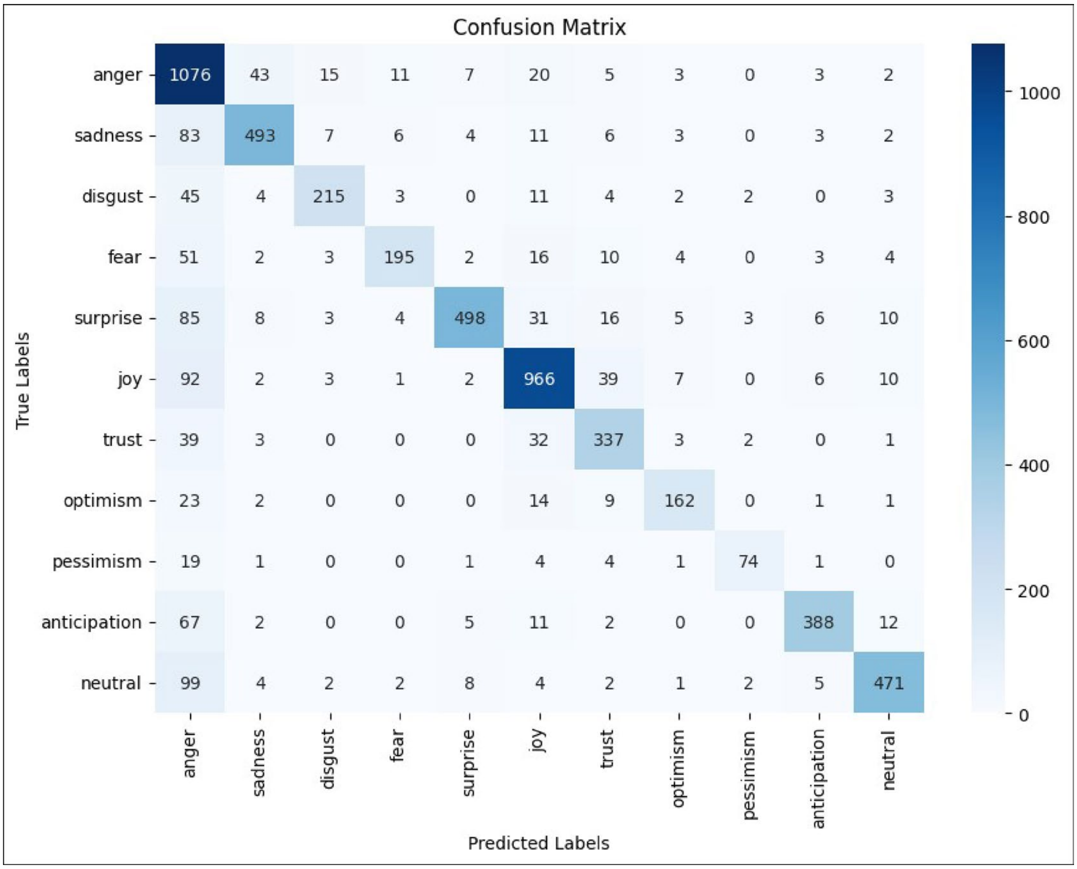


Fig. 9. Best performing ensemble model based on the proposed Hyb-Stack ensemble method.

Authors	Ensemble technique used	Ensemble key features	Performance of ensemble models (F1-score) (%)
[4]	Weighted Average Ensemble	Combines Bangla-BERT-2, XLM-R, Indic-DistilBERT, and Bangla-BERT-1; uses weighted Softmax averaging for prediction; optimized for Bengali emotion texts	80.24
[23]	Weighted Stacked Generalization	Combines BiLSTM and BERT (IndoBERT-large-p2); stacking uses logistic regression as meta-learner, optimized on PRDECT-ID (Indonesia)	75
[33]	Average Voting Ensemble (AVEDL)	Uses BERT, DistilBERT, and RoBERTa; trained on tweets and ERSS call transcripts; supports 5 emotion classes	85.20
[62]	Voting Ensemble Classifier (VEC) with AraBERT embeddings	Combines SGD Classifier, Linear SVC, Multinomial NB, and Ridge Classifier; leverages AraBERT for Arabic sentiment analysis	73.77
[63]	Ensemble of Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM) with AdaBoost	Uses Adaptive Boosting (AdaBoost) to combine predictions from RF, DT, and SVM; optimized for tweet sentiment classification	87.32
[35]	Multi-Level Ensemble (Combines Bagging-based, Boosting-based, and Multinomial Naive Bayes)	Combines Random Forest, Balanced Random Forest, XGBoost, and Naive Bayes models	62
Our work	A new approach—hybrid stack ensemble (Hyb-Stack)	Combines BERT, RoBERTa, DistilBERT, and mBERT using Random Forest as meta-classifier; optimized for multi-label and multilingual emotion classification	90.67

Table 7. Comparison of ensemble learning approaches for emotion classification.

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.

Received: 23 September 2025; Accepted: 29 January 2026
Published online: 07 February 2026

References

1. Yang, S. Emotion detection and analysis techniques based on NLP. 147–153. https://doi.org/10.54254/2755-2721/135/2025.21215 (2025).
2. Merayo, N., Vegas, J., Llamas, C. & Fernández, P. Social network sentiment analysis using hybrid deep learning models. *Appl. Sci.* https://doi.org/10.3390/app132011608 (2023).
3. Thirugnanasammandamoorthi, P., Kumar, H., Ghosh, D., Dhasarathan, C. & Dewangan, R. K. Sentimental analysis and prediction of socioeconomic disasters tweets by ML and regular expression. *J. Intell. Fuzzy Syst.* https://doi.org/10.3233/JIFS-219417 (2024).
4. Das, A., Hoque, M. M., Sharif, O., Dewan, M. A. A. & Siddique, N. TEmoX: Classification of textual emotion using ensemble of transformers. *IEEE Access* **11**(August), 109803–109818. https://doi.org/10.1109/ACCESS.2023.3319455 (2023).
5. Adamu, H. et al. Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning. *Sustainability* https://doi.org/10.3390/su13063497 (2021).
6. Dwarakanath, L., Kamsin, A., Rasheed, R. A., Anandhan, A. & Shuib, L. Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: a review. *IEEE Access* **9**, 68917–68931. https://doi.org/10.1109/ACCESS.2021.3074819 (2021).
7. Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J. & Gaber, M. M. TTL: transformer-based two-phase transfer learning for cross-lingual news event detection. *Int. J. Mach. Learn. Cybern.* https://doi.org/10.1007/s13042-023-01795-9 (2023).
8. Ameer, I. et al. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.* **213**, 118534. https://doi.org/10.1016/j.eswa.2022.118534 (2023).
9. Kowsher, M. et al. Bangla-BERT: Transformer-based efficient model for transfer learning and language understanding. *IEEE Access* **10**(September), 91855–91870. https://doi.org/10.1109/ACCESS.2022.3197662 (2022).
10. Azime I. A. et al. Masakhane-Afrisenti at SemEval-2023 Task 12: Sentiment analysis using afro-centric language models and adapters for low-resource African languages [Online]. Available: http://arxiv.org/abs/2304.06459 (2023).
11. Labib, F. H., Elagamy, M. & Saleh, S. N. EmoBERTa-X: Advanced emotion classifier with multi-head attention and DES for multilabel emotion classification (2025).
12. Kusal, S. et al. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artif. Intell. Rev.* https://doi.org/10.1007/s10462-023-10509-0 (2023).
13. Lak, A. J., Boostani, R., Alenizi, F. A., Mohammed, A. S. & Fakhrahmad, S. M. RoBERTa, ResNeXt and BiLSTM with self-attention: The ultimate trio for customer sentiment analysis. *Appl. Soft Comput.* https://doi.org/10.1016/j.asoc.2024.112018 (2024).
14. Bharti, S. K. et al. Text-based emotion recognition using deep learning approach. *Comput. Intell. Neurosci.* https://doi.org/10.1155/2022/2645381 (2022).
15. Fuad, A. & Al-Yahya, M. Cross-lingual transfer learning for arabic task-oriented dialogue systems using multilingual transformer model mT5. *Mathematics* https://doi.org/10.3390/math10050746 (2022).
16. Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F. & Iglesias, C. A. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **77**, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002 (2017).
17. Satrya, W. F., Aprilliyani, R. & Yossy, E. H. Sentiment analysis of Indonesia police chief using multi-level ensemble model. *Procedia Comput. Sci.* **216**(2022), 620–629. https://doi.org/10.1016/j.procs.2022.12.177 (2022).
18. Du, Y., Wang, Y., Hu, J., Li, X. & Chen, X. An emotion role mining approach based on multiview ensemble learning in social networks. *Inf. Fusion* **88**(July), 100–114. https://doi.org/10.1016/j.inffus.2022.07.010 (2022).
19. Zheng, W. & Jin, M. The effects of class imbalance and training data size on classifier learning: An empirical study. *SN Comput. Sci.* https://doi.org/10.1007/s42979-020-0074-0 (2020).
20. Al Shamsi, A. A. & Abdallah, S. Ensemble stacking model for sentiment analysis of Emirati and Arabic dialects. *J. King Saud Univ. Comput. Inf. Sci.* https://doi.org/10.1016/j.jksuci.2023.101691 (2023).
21. Althobaiti, M. J. Applied sciences Arabic emotion recognition in low-resource settings: A novel diverse model stacking ensemble with self-training (2023).
22. Ghosh, S., Priyankar, A., Ekbal, A. & Bhattacharyya, P. Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. *Knowl. Based Syst.* **260**, 110182. https://doi.org/10.1016/j.knosys.2022.110182 (2023).
23. Pramana, R., Jonathan, M., Yani, H. S. & Sutoyo, R. A comparison of BiLSTM, BERT, and ensemble method for emotion recognition on Indonesia product reviews. *Procedia Comput. Sci.* **245**, 399–408. https://doi.org/10.1016/j.procs.2024.10.266 (2024).
24. Almulihi, A. et al. Ensemble learning based on hybrid deep learning model for heart disease early prediction. *Diagnostics* **12**(12), 1–17. https://doi.org/10.3390/diagnostics12123215 (2022).
25. Ghorbanali, A., Sohrabi, M. K. & Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Inf. Process. Manag.* **59**(3), 102929. https://doi.org/10.1016/j.ipm.2022.102929 (2022).
26. Iqbal, M. A., Das, A., Sharif, O., Hoque, M. M. & Sarker, I. H. BEmoC: A corpus for identifying emotion in Bengali texts. *SN Comput. Sci.* **3**(2), 1–17. https://doi.org/10.1007/s42979-022-01028-w (2022).
27. Plutchik, R. *Emotions and Life: Perspectives from Psychology, Biology, and Evolution* (American Psychological Association, 2003).
28. Alslaity, A. & Orji, R. Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behav. Inf. Technol.* **43**(1), 139–164. https://doi.org/10.1080/0144929X.2022.2156387 (2024).
29. Alsmadi, I. et al. Adversarial machine learning in text processing: A literature survey. *IEEE Access* **10**(January), 17043–17077. https://doi.org/10.1109/ACCESS.2022.3146405 (2022).
30. de Oliveira R. F. et al. Enhanced Reader.pdf (2018).
31. Behzadidoost, R., Mahan, F. & Izadkhah, H. Granular computing-based deep learning for text classification. *Inf. Sci. (NY)* **652**, 119746. https://doi.org/10.1016/j.ins.2023.119746 (2024).
32. Meng, Q. et al. Electric power audit text classification with multi-grained pre-trained language model. *IEEE Access* **11**, 13510–13518. https://doi.org/10.1109/ACCESS.2023.3240162 (2023).
33. Nimmi, K., Janet, B., Selvan, A. K. & Sivakumaran, N. Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset. *Appl. Soft Comput.* https://doi.org/10.1016/j.asoc.2022.108842 (2022).
34. Shoubaki, H., Abdallah, S. & Shaalan, K. Deep learning techniques for identifying poets in Arabic Poetry: A focus on LSTM and Bi-LSTM. *Procedia Comput. Sci.* **244**, 461–470. https://doi.org/10.1016/j.procs.2024.10.221 (2024).
35. Malik, A., Behera, D. K., Hota, J. & Swain, A. R. Ensemble graph neural networks for fake news detection using user engagement and text features. *Results Eng.* https://doi.org/10.1016/j.rineng.2024.103081 (2024).
36. Alrasheedy, M. N., Muniyandi, R. C. & Fauzi, F. Text-based emotion detection and applications: A literature review. *Int. Conf. Cyber Resilience ICCR* https://doi.org/10.1109/ICCR56254.2022.9995902 (2022).
37. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019—2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 4171–4186 (2019).
38. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter, 2–6 (2019).
39. Pradnyana, G. A., Anggraeni, W., Yuniarno, E. M. & Purnomo, M. H. An explainable ensemble model for revealing the level of depression in social media by considering personality traits and sentiment polarity pattern. *Online Soc. Netw. Media* https://doi.org/10.1016/j.osnem.2025.100307 (2025).
40. Bogatinovski, J., Todorovski, L., Džeroski, S. & Kocev, D. Comprehensive comparative study of multi-label classification methods. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2022.117215 (2022).

41. Wu, G., Zheng, R., Tian, Y. & Liu, D. Joint ranking SVM and binary relevance with robust low-rank learning for multi-label classification. *Neural Netw.* **122**, 24–39. https://doi.org/10.1016/j.neunet.2019.10.002 (2020).
42. Tsoumakas, G., Katakis, I. & Vlahavas, I. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* **23**(7), 1079–1089. https://doi.org/10.1109/TKDE.2010.164 (2011).
43. Khan, T., Yasir, M. & Choi, C. Attention-enhanced optimized deep ensemble network for effective facial emotion recognition. *Alex. Eng. J.* **119**, 111–123. https://doi.org/10.1016/j.aej.2025.01.078 (2025).
44. Saleh, S. N. Enhancing multilabel classification for unbalanced COVID-19 vaccination hesitancy tweets using ensemble learning. *Comput. Biol. Med.* **184**(June), 2025. https://doi.org/10.1016/j.compbiomed.2024.109437 (2024).
45. Garcia-Diaz, J. A., Garcia-Sanchez, F. & Valencia-Garcia, R. Smart analysis of economics sentiment in Spanish based on linguistic features and transformers. *IEEE Access* **11**(February), 14211–14224. https://doi.org/10.1109/ACCESS.2023.3244065 (2023).
46. Kratzwald, B. & Feuerriegel, S. Putting question-answering systems into practice: Transfer learning for efficient domain customization. *ACM Trans. Manag. Inf. Syst.* **9**(4), 1–17. https://doi.org/10.1145/3309706 (2019).
47. Delbrouck, J. B., Tits, N., Brousmiche, M. & Dupont, S. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *Proc. Annu. Meet. Assoc. Comput. Linguist.* https://doi.org/10.18653/v1/2020.challengehml-1.1 (2020).
48. Huan, J. L., Sekh, A. A., Quek, C. & Prasad, D. K. Emotionally charged text classification with deep learning and sentiment semantic. *Neural Comput. Appl.* **34**(3), 2341–2351. https://doi.org/10.1007/s00521-021-06542-1 (2022).
49. Mohammadi, A. & Shaverizade, A. Ensemble deep learning for aspect-based sentiment analysis. *Int. J. Nonlinear Anal. Appl.* **12**, 29–38 (2021).
50. Nguyen-The, M., Lamghari, S., Bilodeau, G. A. & Rockemann, J. Leveraging sentiment analysis knowledge to solve emotion detection tasks. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 13643 LNCS, 405–416. https://doi.org/10.1007/978-3-031-37660-3_29 (2023).
51. Frye, R. H. & Wilson, D. C. Comparative analysis of transformers to support fine-grained emotion detection in short-text data. *Proc. Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS* https://doi.org/10.32473/flairs.v35i.130612 (2022).
52. Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E. & Mohammad, R. An Ensemble-learning-based technique for bimodal sentiment analysis. *Big Data Cogn. Comput.* https://doi.org/10.3390/bdcc7020085 (2023).
53. Han, T. et al. Text emotion recognition based on XLNet-BiGRU-Att. *Electronics* **12**(12), 1–15. https://doi.org/10.3390/electronics12122704 (2023).
54. Kamran, S. et al. EmoDNN: Understanding emotions from short texts through a deep neural network ensemble. *Neural Comput. Appl.* **35**(18), 13565–13582. https://doi.org/10.1007/s00521-023-08435-x (2023).
55. Lee, S. J., Lim, J. Y., Paas, L. & Ahn, H. S. Transformer transfer learning emotion detection model: Synchronizing socially agreed and self-reported emotions in big data. *Neural Comput. Appl.* **35**(15), 10945–10956. https://doi.org/10.1007/s00521-023-08276-8 (2023).
56. Zhen, H., Shang, W. & Zhang, W. Sentiment analysis of hybrid network model based on attention. *Int. J. Softw. Innov.* **11**(1), 1–17. https://doi.org/10.4018/IJSI.327364 (2023).
57. Li, Y., Zhu, R. & Li, W. CorMulT: A semi-supervised modality correlation-aware multimodal transformer for sentiment analysis. *IEEE Trans. Affect. Comput.* https://doi.org/10.1109/TAFFC.2025.3559866 (2025).
58. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5999–6009 (2017).
59. Liu, Y. et al. RoBERTa: A robustly optimized BERT pretraining approach, No. 1 (2019).
60. Mohammad, S. M. & Bravo-marquez, F. SemEval-2018 task 1: Affect in Tweets, 1–17 (2018).
61. Ibrohim, M. O. Multi-label hate speech and abusive language detection in Indonesia Twitter, 46–57 (2019).
62. Ghoul, D., Patrix, J., Lejeune, G. & Verny, J. A combined AraBERT and Voting Ensemble classifier model for Arabic sentiment analysis. *Nat. Lang. Process. J.* **8**, 100100. https://doi.org/10.1016/j.nlp.2024.100100 (2024).
63. Kp, V., Ab, R., Hl, G., Ravi, V. & Krichen, M. A tweet sentiment classification approach using an ensemble classifier. *Int. J. Cogn. Comput. Eng.* **5**, 170–177. https://doi.org/10.1016/j.ijcce.2024.04.001 (2024).

Acknowledgements

The first author gratefully acknowledges the support of the Petroleum Technology Development Fund (PTDF), Abuja, Nigeria, for sponsoring his PhD studies, which enabled him to carry out this research 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 and editing, supervision. Nurul Amelina Nasharuddin: Conceptualization, writing—review and editing, supervision.

Funding

This research was funded by Universiti Putra Malaysia.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-38172-9.

Correspondence and requests for materials should be addressed to H.A. or M.A.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.