big data and cognitive computing

MDPI

Article

An Attention-Based BERT–CNN–BiLSTM Model for Depression Detection from Emojis in Social Media Text

Joel Philip Thekkekara * and Sira Yongchareon

School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology; Auckland 1010, New Zealand
* Correspondence: tjoephd@gmail.com

Abstract

Depression represents a critical global mental health challenge, with social media offering unprecedented opportunities for early detection through computational analysis. We propose a novel BERT–CNN–BiLSTM architecture with attention mechanisms that systematically integrate emoji usage patterns—fundamental components of digital emotional expression overlooked by existing approaches. Evaluated on the SuicidEmoji dataset, our model achieves 97.12% accuracy, 94.56% precision, 93.44% F1-score, 85.67% MCC, and 91.23% AUC-ROC. Analysis reveals distinct emoji patterns: depressed users favour negative emojis (😔 13.9%, 😢 12.8%, 💔 6.7%) while controls prefer positive expressions (😂 16.5%, 😊 11.0%, 😎 10.2%). The attention mechanism identifies key linguistic markers, including emotional indicators, personal pronouns, and emoji features, providing interpretable insights into depression-related language. Our findings suggest that the integration of emojis substantially improves optimal social media-based mental health detection systems.

Keywords: depression detection; natural language processing; deep learning; emoji analysis; mental health

Academic Editor: Ximing Li

Received: 25 September 2025
Revised: 20 November 2025
Accepted: 25 November 2025
Published: 3 December 2025

Citation: Thekkekara, J.P.; Yongchareon, S. An Attention-Based BERT–CNN–BiLSTM Model for Depression Detection from Emojis in Social Media Text. *Big Data Cogn. Comput.* **2025**, *9*, 310. https://doi.org/10.3390/bdcc9120310

1. Introduction

Depression affects approximately 264 million individuals worldwide, with nearly 50% not receiving adequate professional care due to stigma and limited access to mental health services [1,2]. Traditional diagnostic approaches, relying on clinical assessments and standardised questionnaires, face substantial limitations, including high false-positive rates and dependence on patient self-reporting, which is influenced by social desirability bias [3]. Multi-stage screening strategies have been recommended to address these challenges; however, their implementation remains complex and resource-intensive [4]. The widespread adoption of social media platforms has created unprecedented opportunities for mental health analysis [5]. Users increasingly share personal struggles and mental health challenges through user-generated content [6], with machine-learning techniques used for timely identification and intervention [7,8].

Digital communication has transformed, with over 3600 standardised emojis serving as emotional communicators across linguistic barriers [9]. Apple's 2024 introduction of Genmoji (AI-generated, context-specific emojis) further expands personalised emotional expression, a domain that is exponentially growing. Research demonstrates that emojis reduce communication ambiguity by 40% and increase emotional expressiveness by 65% compared to text-only messages [10]. However, this shift in visual emotional expression, particularly

among younger demographics [11], challenges traditional text-based depression detection models, which cannot interpret the complex emotional semantics embedded in emoji usage patterns [12,13].

Current depression detection models overlook emoji features, either removing them during preprocessing or treating them peripherally, thereby discarding emotional information which constitutes 15% of mental health-related social media content. Despite computational linguistics research demonstrating emoji patterns as reliable psychological indicators, existing approaches fail to systematically integrate emoji sentiment and usage patterns as primary classification features [14,15]. This limitation is critical, given that depressed individuals exhibit distinct emoji characteristics. Studies show that individuals with poorer mental health tend to prefer negative emojis [16], while those with higher depression and anxiety express negative emotions more frequently on social media. Additionally, lower use of positive emoticons correlates with increased emotional distress [17]. These findings suggest that emoji patterns may serve as meaningful indicators of underlying emotional states.

This study addresses these limitations through a novel attention-based BERT–CNN–BiLSTM architecture that elevates emoji features from auxiliary to primary indicators in the detection of depression. Our methodological approach systematically integrates emoji sentiment information, usage patterns and contextual relationships within the textual analysis framework, treating visual emotional indicators as semantically equivalent to linguistic content rather than supplementary features. The model achieves strong performance across multiple evaluation metrics, including MCC on the SuicidEmoji dataset, with attention mechanism visualisation revealing the relative importance of textual and emoji features in classification decisions. Whilst previous work [15] employed similar architecture, that study removed emojis during preprocessing. The current study extends this approach by systematically preserving emoji information through descriptive text conversion, achieving +1.89% accuracy improvement and providing the first analysis of emoji distribution patterns in depression-related content.

The primary contribution of this work is to present a systematic investigation of emoji usage patterns, sentiment distributions and temporal trends for the first time in depression-related social media text data through an attention-based BERT–CNN–BiLSTM model that achieves state-of-the-art results across multiple evaluation metrics including MCC, benchmarking deep-learning models for depression detection on the SuicidEmoji dataset and provide attention mechanism visualisation revealing the relative importance of textual and emoji features in classification decisions.

The remainder of this paper is organised as follows: Section 2 reviews related work in depression detection and emoji analysis; Section 3 details the proposed attention-based BERT–CNN–BiLSTM (BCBA) architecture and feature integration approach; Section 4 presents the dataset and preprocessing methodology; the experimental results and comparison with baseline models; as well as providing a discussion of findings and attention visualisation analysis and thereafter, concludes with implications and future research directions.

2. Related Work and Background

The systematic analysis of emojis for computational applications has emerged as a critical research area for mental health detection. This section reviews existing approaches to emoji-based sentiment analysis, their methodological contributions, and identified limitations that motivate our proposed framework. Initial efforts in emoji sentiment analysis focused on establishing quantitative measures for emotional valence [18]. This effort developed sentiment scores for 751 emojis through crowdsourced annotations (n = 83,000), creating the first comprehensive emoji–emotion mapping. Their annotation framework

achieved an inter-rater reliability of k = 0.72 but exhibited significant cultural bias, with 89% of annotators from Western countries. This foundational work established the feasibility of quantifying emoji sentiment but highlighted the challenge of cross-cultural validity.

Recent advances have explored emoji representation strategies within deep-learning architectures. A comparative analysis of symbolic versus textual emoji representations using 77,439 domain-specific tweets was conducted by [19]. Their transformer-based classifier demonstrated that textual emoji descriptions improved F1-scores by 3.2% over symbolic representations (0.847 vs. 0.821). However, the evaluation remained constrained to English-language data and a single domain, limiting the generalisability of representation strategies across linguistic and topical boundaries. Personality-conditioned emoji generation using Parameter-Efficient Fine-Tuning (PEFT) was explored, achieving 87.3% relevance scores for generated emojis [20]. Their approach demonstrated that personality embeddings could modulate emoji selection in language models, though generalisation remained limited to English-language models and required personality annotations unavailable in most mental health datasets.

Addressing data quality issues in social media, ref. [21] developed a three-stage reverse-engineering methodology for recovering emojis from corrupted text encodings. Applied to 76,914 tweets, their approach successfully reconstructed 157,748 emoji instances with 92.1% precision, improving downstream readability metrics by 34.7%. The method's reliance on UTF-8/UTF-16 encoding patterns; however, constrains applicability to platforms with proprietary emoji implementations. An architecture, integrating Dominance Guiding Defense Optimisation with bidirectional recurrent networks (DGDO–BiLSTM) was proposed by [22]. Their model achieved 89.4% accuracy on multilingual sentiment detection tasks, outperforming standard BiLSTM by 7.2 percentage points. The optimisation strategy specifically targeted emoji–text interactions through attention weights, though evaluation was limited to high-resource languages with established emoji usage patterns.

The complexity of emoji interpretation in context has motivated several computational approaches. Emoji prediction as a multi-class classification task for SemEval-2018 was formulated by [23], establishing benchmarks with 47.1% top-1 accuracy across 20 emoji classes. Their analysis revealed that contextual embeddings improved prediction accuracy by 12.3% over bag-of-words baselines, demonstrating the importance of sequential modelling. This work was further complemented by [10], quantifying interpretation variance, finding agreement rates as low as 25.4% across demographic groups for ambiguous emojis, with significant variations across age groups and cultural backgrounds affecting emoji-based mental health analysis reliability.

The application of emoji patterns for psychological profiling has demonstrated promising results [24]. Employing LSTM networks, researchers classified personality traits from emoji usage frequencies, achieving 95.48% accuracy across 16 personality categories (encompassing four psychological dimensions: introversion–extroversion, happiness–depression, optimism–pessimism, and neuroticism–calmness) derived from 13,688 WhatsApp conversations. The model extracted 128-dimensional emoji embeddings trained on usage co-occurrences; however, the computational requirements for image-based processing limited the feasibility of real-time deployment [25]. Pursuing unsupervised emotion discovery, they applied their EmDMM framework, which utilises Dirichlet Multinomial Mixture modelling for emoji clusters. Analysis of COVID-19 related tweets yielded eight distinct emotional categories with 76.3% cluster purity. While effective for dominant emotion identification during crisis situations, the approach struggled with mixed-valence expressions containing contradictory emoji combinations, achieving only 52.1% accuracy on multi-emotion texts. Recent architectures have incorporated specialised attention mechanisms for emoji-aware sentiment analysis [26]. Proposed dual-modality attention networks jointly process textual

and emoji inputs, achieving 91.7% accuracy on educational feedback datasets. The attention mechanism learnt to weight emoji contributions dynamically, with emoji features accounting for 23.4% of final predictions on average. The model's evaluation remained limited to domain-specific corpora, with unverified generalisability across different languages and informal communication contexts.

These advances collectively demonstrate that emoji-based features provide valuable psychological insights that extend beyond simple sentiment analysis, offering opportunities for comprehensive mental health assessment through digital communication patterns which form the premises for the work exhibited in this research.

3. Architectural Model

Our proposed architecture (Figure 1) integrates multiple deep-learning components to leverage semantic understanding, local pattern recognition, sequential modelling and attention-based feature weighting. The model design reflects the principle that emoji-enhanced depression detection requires comprehensive analysis of both textual semantics and visual emotional indicators.

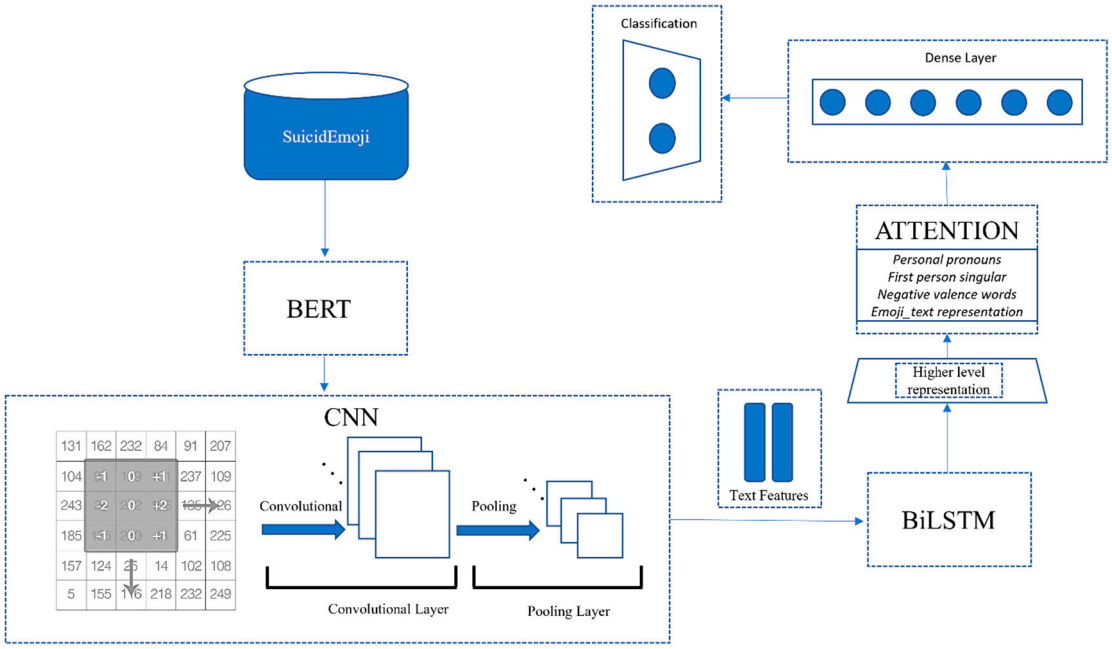

Figure 1. BERT–CNN–BiLSTM–ATTN architecture for depression classification using emojis in text.

3.1. BERT–CNN–BiLSTM Attention (BCBA) Model

Each component in our BCBA architecture addresses distinct aspects of emoji-enhanced depression detection. The BERT layer provides contextual embeddings that capture semantic relationships between emoji descriptors and surrounding text. By converting emojis to descriptive text (e.g., 😔 → ":pensive face:"), we enable BERT's pre-trained linguistic knowledge to process emoji semantics as integral textual elements rather than isolated symbols. This approach leverages BERT's understanding of descriptive language whilst maintaining the emotional information encoded in emoji usage.

The CNN component extracts local n-gram patterns that indicate emotional intensity. Multiple filter sizes (3, 4, 5) capture varying phrase lengths, such as "can't crying face take" and "feel pensive face worthless," which represent characteristic depressive language structures. These local patterns complement BERT's broader contextual understanding by identifying position-invariant emotional markers that may appear at different loca-

tions within posts. The max-pooling operation ensures that salient emotional patterns are captured, regardless of their position in the sequence.

The BiLSTM component models sequential dependencies in both forward and backward directions, capturing the temporal flow of emotional expression across posts. This bidirectional processing is particularly important as depressive language often exhibits progression patterns—for instance, initial attempts at positivity, followed by expressions of hopelessness. The BiLSTM's capacity to model long-range dependencies enables architecture to understand how emotional tone evolves throughout a post.

The attention mechanism weights the importance of different features, allowing the model to focus on emotionally significant tokens, including personal pronouns ("I", "me"), negative valence words ("hopeless", "worthless"), negations ("never", "can't") and emoji–text descriptors. This provides both performance enhancement and interpretability. Our attention visualisations demonstrate that the model learns to focus on clinically relevant linguistic markers associated with depression. The attention weights reveal which features contribute most strongly to classification decisions, offering transparency often lacking in deep-learning approaches to mental health detection.

We start by embedding the input text using BERT. This allows the model to capture the contextual meaning of the text and the emoji–text description. The BERT model outputs a sequence of embeddings for each token in the text. The BERT layer generates a matrix of size (T,d) where T is the number of tokens in the text and d is the embedding dimension.

$$\mathrm{H_{BERT}} = \mathrm{BERT}(text)$$

The BERT layer is fine-tuned during training to adapt representations specifically for depression detection tasks, while preserving general linguistic knowledge, with a hidden size of 768 dimensions, a maximum sequence length of 250 tokens, and a dropout rate of 0.1, using the uncased model with 12 transformer layers. After obtaining the BERT embeddings, we pass them through one-dimensional convolutional layers. CNNs are used to capture local features or patterns within the text, such as repeated phrases or emotive cues like "I feel worthless crying face loudly crying face" or "I can't do this crying face".

$$F_i = ReLU(W_i \ * \ H_{BERT} \ + \ b_i)$$

Here, $*$ denotes the convolution operation, W_i are the filter weights and b_i are the biases. The output $\mathrm{F_i}$ represents the filtered feature maps that highlight important local patterns. We apply multiple filter sizes to capture different granularities of local features and thereafter, max pooling to concatenate features from all filter sizes. The CNN component employs multiple filter sizes of [3–5] to capture n-gram patterns of varying lengths, with 128 filters per size yielding a total of 384 filters. Each convolutional operation is followed by ReLU activation for non-linear transformation, and 1-max pooling is applied to extract the most salient features from each filter's output, effectively identifying the most important patterns regardless of their position in the input sequence.

The CNN-generated features are passed into a BiLSTM layer. BiLSTM is used to capture long-range sequential dependencies and temporal patterns in both forward ($\mathrm{LSTM_f}$) and backward ($\mathrm{LSTM_b}$) directions as depressive language may evolve over a sequence of sentences (e.g., a conversation or a series of posts).

$$\overrightarrow{\mathrm{h}_t} = \mathrm{LSTM_f}(\mathrm{F}_t)$$

$$\overleftarrow{\mathrm{h}_t} = \mathrm{LSTM_b}(\mathrm{F}_t)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

This component is essential for understanding the sequential flow of emotional expressions and linguistic patterns. To enhance the model's ability to focus on critical tokens or phrases (e.g., "hopeless crying face", "I can't take it loudly crying face"), a multi-head attention mechanism is applied after the BiLSTM. The attention mechanism computes a weight for each token based on its relevance to the task, allowing the model to focus on emotionally significant words such as personal pronouns, first-person singular and negative valence words, as well as the associated emoji–text, which provides further context to the given text. This could provide further interpretability for future clinical applications.

$$\alpha_t = \frac{\exp\left(v^\top \tanh(W_a h_t + b_a)\right)}{\sum_{t'=1}^{T} \exp\left(v^\top \tanh(W_a h_{t'} + b_a)\right)}$$

Here, v is a weight vector, W_a is a matrix of learnable weights, and b_a is the bias term. This results in a weighted sum of the hidden states:

$$h_{att} = \sum_{t=1}^{T} \alpha_t h_t$$

The emoji embeddings, obtained from the previous preprocessing step, are concatenated with the output from the attention mechanism. This ensures that emotional cues from emojis (e.g., 😞 for sadness) are taken into account in the final classification decision.

$$h_{final} = \left[h_{att}; E_{emoji}\right]$$

The final output h_{final} is passed through a fully connected dense layer, followed by a SoftMax activation function for binary classification (e.g., 0 for non-depressed and 1 for depressed).

$$\hat{y} = \text{softmax}(W_c h_{final} + b_c)$$

where $\hat{y}$ is the predicted label for depression, and W_c and b_c are the weights and biases of the final classification layer.

3.2. Experiments and Evaluation:

This section provides an overview of the dataset utilised, the experimental setup for the various models tested and the evaluation metrics applied. Given the ongoing challenges in the medical domain regarding the classification of imbalanced datasets, this study retains the original class distribution of the SuicidEmoji dataset, i.e., its imbalanced nature, to assess the performance of several deep-learning models, including the proposed architecture. The deep-learning models within this architecture utilise the ReLU activation function for the hidden layers and the SoftMax function for the output layer. The loss function is a critical component, as it determines how effectively the output layer is integrated with the rest of the network. In our experiments, given the binary classification task, binary cross-entropy was selected as the loss function.

3.3. Experimental Setup

The input processing pipeline handles both textual content and emoji features through a unified tokenisation approach. Given an input sequence S = [w_1, w_2, …, w_n, e_1, e_2, …., e_m], where w◎ represents words and e◎ represents emojis, the preprocessing steps such as text normalisation, which includes converting to lowercase, removing URLs and user mentions, and emoji preservation, which maintains the original Unicode representation

of emojis, as well as tokenisation, which uses the BERT tokeniser with vocabulary size of 30,522 tokens, as well as sequence padding to standardise input length to 250 tokens based on dataset statistics.

3.4. Dataset and Setup

3.4.1. Dataset Details

In this study, we utilise the SuicidEmoji dataset [27], a comprehensive collection of social media posts designed to aid in the research and detection of suicide-related content using emojis. This dataset is specifically derived from a variety of publicly available sources and aims to address the increasing need for automated systems to identify potentially harmful content related to suicide or self-harm on social media platforms. The dataset comprises textual data with emojis from two Reddit corpora collected via Pushshift API: (i) SuicideReddit, containing /SuicideWatch posts (December 2008–January 2021) with /teenagers as control group; and (ii) Robin, combining /SuicideWatch with 13 other subreddits (e.g., /CasualConversation, /self, /TIFU) from 2019 onwards. While posts lack manual annotation, self-labelling within /SuicideWatch is considered reliable given the data scale and prior validation studies [27]. After eliminating duplicates, the researchers obtained approximately 1.3 million posts across both datasets. For the purpose of analysis, they have focused only on posts containing emojis. To decode the emoji characters, we utilised the demoji package, as most emojis are encoded in the Unicode standard. The final dataset consists of 25,051 emoji instances, including 2329 posts related to suicide and 22,722 posts from the control group as illustrated in Table 1.

Table 1. Detailed statistics of the SuicidEmoji dataset.

Metric	Suicide	Control
# of posts	2329	22,722
Avg. length	225.82	123.94
Avg. # of emojis	2.60	10.14
Avg. # of different emojis	1.37	2.30
Top 10 emojis	😭 😃 ❤️ 😔 😂 😢 😞 💔 💩 😊	🤡 🗿 🦀 ⏰ 😂 🇺🇸 💰 😎 🅱️ 😭
Top 10 emojis (rm dup.)	❤️ 😔 😭 😞 😢 😂 💔 🖤 ✌️ 😊	😎 😂 😔 😳 😭 ❤️ 😩 😊 😤 🖤
Emoji sentiment scores	0.66, −0.15, −0.09, −0.15, 0.01, 0.22, −0.12, 0.75, 0.46, 0.64	0.49, 0.22, −0.15, 0.02, −0.09, 0.66, −0.37, 0.64, null, 0.75
Top 5 co-used emojis	👋, 😔 | 💔, 😭 | 😢, 😭 | 💔, 😔 | 😔, 😢	👈, 👉 | 💦, 😩 | 👉, 😎 | 😂, 🤣 | 👁️, 👄

A few limitations warrant acknowledgement regarding the SuicidEmoji dataset. Firstly, classification relies on subreddit membership (r/SuicideWatch vs. control subreddits) rather than clinical diagnoses, introducing potential self-labelling bias. Whilst [27] validated this approach through analysis of characteristic linguistic patterns in their data collection, not all r/SuicideWatch participants necessarily meet clinical criteria for depression, and control groups may contain undiagnosed individuals. Secondly, Reddit's demographics introduce systematic biases: users skew younger (predominantly 18–29 years), male (approximately 62%) and towards Western English-speaking populations. Thirdly, r/SuicideWatch represents individuals actively seeking support online, which may not reflect isolated or withdrawn individuals who are less likely to engage in digital mental health communities. These limitations could mean our model may detect "Text-style expressions of depression (with emojis) in support-seeking individuals" (e.g., text subreddit conven-

tions, community-specific emoji usage) rather than universal depression markers applicable across all populations and contexts.

3.4.2. Training Details

$$Loss = -\frac{1}{\text{outputsize}} \sum_{i=1}^{\text{outputsize}} y_i \cdot \log\hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

where $\hat{y}_i$ is the ith scalar value in the model output, y_i is the corresponding target value, and output size is the number of scalar values in the model output.

3.4.3. Evaluation Metric

In this study, we evaluate classification performance using accuracy, precision, F1 score, AUC-ROC, and MCC. Since precision, recall, and F1 score ignore true negatives (TN), while accuracy suffers from class imbalance sensitivity, we prioritise MCC—a correlation coefficient that evaluates both classes. This is critical for depression detection, where correctly identifying depressed individuals and avoiding false positives are equally important. MCC provides a superior classification assessment by incorporating all elements of the confusion matrix [15].

$$Accuracy = \frac{TP + TN}{TP + TN} \qquad Precision = \frac{TP}{TP + FP}$$

$$F1\ Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \qquad MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC ranges from -1 to 1, where MCC = 1 when FP = FN = 0 (perfect classification) and MCC = -1 when TP = TN = 0 (complete misclassification). As a symmetric measure utilising all confusion matrix elements, MCC provides an unbiased evaluation regardless of class distribution, making it particularly suitable for imbalanced datasets, where higher values indicate an accurate prediction of both classes.

4. Results and Discussions

This work presents the first comprehensive analysis of emoji usage patterns for understanding emotional expression and benchmarks multiple deep-learning models for depression detection using the SuicidEmoji task (Table 2).

Table 2. Benchmarking SuicidEmoji.

Model	Accuracy	F1-Score	AUC-ROC	Precision	MCC
LSTM	94.26%	0.81	0.76	0.90	0.65
BiLSTM	94.30%	0.81	0.76	0.90	0.65
CNN	95.41%	0.86	0.82	0.87	0.74
CNN–LSTM	95.39%	0.86	0.81	0.91	0.73
CNN–BiLSTM	95.47%	0.86	0.82	0.90	0.72
BiLSTM–Attention	93.75%	0.82	0.79	0.84	0.64
CNN–BiGRU	93.30%	0.80	0.75	0.88	0.62
BERT–CNN–BiLSTM–Attention (BCBA)	97.12%	0.94	0.91	0.95	0.86

To validate the performance superiority of BCBA, we conducted pairwise comparisons using paired t-tests ($\alpha = 0.05$). Results demonstrate statistically significant differences:

The Wilcoxon signed-rank test confirmed that BCBA achieved the highest mean ranking (8.00) amongst all models, with all pairwise comparisons favouring BCBA ($p < 0.05$). These results confirm that performance improvements are statistically significant, rather than arising from random variation.

Paired *t*-test analysis [28] ($\alpha = 0.05$) revealed statistically equivalent performance between LSTM–BiLSTM ($p = 0.13$), CNN–BiLSTM–CNN–LSTM ($p = 0.066$), and CNN–CNN–BiLSTM ($p = 0.129$) pairs. All other pairwise comparisons, particularly involving the proposed BCBA architecture, showed significant differences ($p < 0.05$). CNN-based architectures demonstrated superior performance across experiments, except CNN–BiGRU.

Statistical validation of model performance was conducted using paired *t*-test analysis, following established methodologies [28], with a significance threshold of $\alpha = 0.05$. The comparative analysis revealed that several model pairs exhibited statistically equivalent performance: LSTM versus BiLSTM architectures ($p = 0.13$), CNN–BiLSTM compared to CNN–LSTM ($p = 0.066$), and CNN against CNN–BiLSTM ($p = 0.129$). Conversely, all remaining pairwise comparisons, particularly those involving our proposed BCBA architecture, demonstrated statistically significant performance differences ($p < 0.05$). Our empirical findings indicate that CNN-based architectures outperform alternative approaches, with the notable exception of the CNN–BiGRU configuration.

This observed limitation of the CNN–BiGRU model may be attributed to the theoretical superiority of LSTM units over GRU components in capturing extended temporal dependencies. Previous research has established that LSTM architectures demonstrate an enhanced capacity for modelling long-range sequential relationships compared to their GRU counterparts [29,30]. The Wilcoxon signed-rank test [31] was subsequently employed as a non-parametric substitute for the paired *t*-test to evaluate two related samples. This analysis demonstrated optimal performance for the BCBA model relative to all competing approaches, with negative ranks consistently favouring BCBA and achieving the highest mean rank score.

The Friedman test ($p < 0.05$) revealed significant performance differences between models, with the BCBA model achieving the highest mean ranking (8.00). This superiority was corroborated through paired *t*-tests and Wilcoxon tests, with 10-fold cross-validation results (Table 3) demonstrating consistent outperformance. However, accuracy metrics alone proved inadequate for imbalanced datasets [32]. Model evaluation optimised AUC-ROC and MCC metrics as both comprehensively optimises all confusion matrix components without class bias. AUC-ROC quantifies the discriminative capacity between classes [31], while MCC provides a robust evaluation for imbalanced datasets by incorporating true negatives that are typically overlooked by other metrics. Table 2 confirms that higher AUC-ROC and MCC scores indicate superior performance, whereas models achieving >90% accuracy may still perform poorly on imbalanced data—a critical consideration for depression detection applications. Considering the metrics outlined in Section 4, the BCBA architecture delivered optimal results amongst all experimental models, achieving the highest AUC-ROC, MCC, and accuracy scores. Whilst BERT and CNN layers contributed meaningfully, BiLSTM and Attention layers provided improvements in overall performance and word- and emoji–text-level contribution insights for classification decisions. For example, attention heatmap optimising showed word-level importance in classification decisions (Figures 2 and 3). Darker colours indicate higher attention weights, which includes the emoji–text replacement. The model successfully focuses on emotional indicators ("depressed", "hopeless", " 💔 ") for depressed posts (Figure 2) and positive sentiment words ("wonderful", " 😊 ", "great") for control posts (Figure 3), demonstrating learned semantic understanding, in addition to the theory that linguistic dimensions like personal pronouns (I, them, her), first-person singular (I, me, mine), and negations (no, not, never)

have a great correlation with depression [8]. The emoji integration impact analysis (Table 4) demonstrates the methodology involving the transformation of emojis into descriptive textual representations, which are subsequently integrated with original content, yielding superior performance outcomes.

Table 3. Statistical significance analysis.

Model Comparison	Observed Difference	Significance
BCBA vs. CNN–BiLSTM	+1.65% accuracy, +0.08 F1, +0.14 MCC	$p < 0.01$
BCBA vs. BiLSTM–Attention	+3.37% accuracy, +0.12 F1, +0.22 MCC	$p < 0.001$
BCBA vs. CNN–LSTM	+1.73% accuracy, +0.08 F1, +0.13 MCC	$p < 0.01$
BCBA vs. CNN	+1.71% accuracy, +0.08 F1, +0.12 MCC	$p < 0.05$

Figure 2. Attention visualisation (depressed).

Figure 3. Attention visualisation (non-depressed).

Table 4. 10-fold cross-validation statistics.

Metric	Mean	Std Dev	95% CI	Min	Max
Accuracy	96.89%	0.34%	[96.61%, 97.17%]	96.45%	97.32%
Precision	94.23%	0.67%	[93.72%, 94.74%]	93.12%	95.01%
F1-Score	93.12%	0.45%	[92.78%, 93.46%]	92.56%	93.78%
MCC	85.34%	0.78%	[84.74%, 85.94%]	84.23%	86.45%
AUC-ROC	90.98%	0.43%	[90.64%, 91.32%]	90.34%	91.67%

This approach exhibited advantages through several mechanisms. Firstly, the conversion facilitated seamless compatibility with BERT architecture, as emoji descriptions were processed as conventional text tokens, enabling comprehensive exploitation of BERT's contextual comprehension capabilities. Secondly, the descriptive transformation preserved semantic integrity by maintaining the emotional significance of emojis while ensuring accessibility for transformer-based architectures. Thirdly, emoji descriptions demonstrated natural integration with surrounding textual content, thereby optimising attention mechanism functionality through enhanced contextual coherence. Finally, the descriptive representations contributed to increased informational density by incorporating emotional context, which substantially enriched the overall textual representation quality. For example, the original text such as "I feel so depressed and hopeless 😢 nothing matters anymore" becomes "I feel so depressed and hopeless crying face nothing matters anymore" following emoji conversion. This modification yields informational gains through the explicit addition of "crying face", which provides direct emotional validation of the expressed textual sentiment.

The observed performance improvements, including a 1.89% accuracy enhancement and a 0.05 F1-score increase (Table 5), substantiate that the emoji-to-text conversion methodology delivers substantial benefits by augmenting textual content with emotional intelligence, rather than treating emojis as discrete feature components. This approach demonstrates that enriching textual representations through emoji integration significantly enhances model performance in sentiment analysis tasks.

Table 5. Emoji integration impact analysis.

Configuration	Accuracy	F1-Score	Performance Change
Text only	95.23%	0.88	Baseline
Text + Emoji (descriptive)	97.12%	0.93	+1.89%/+0.05

Our comprehensive emoji analysis provides profound insights into digital emotional expression patterns, which have significant implications for mental health detection. Emoji utilisation patterns revealed significant cultural and temporal dimensions, enhancing digital mental health expression understanding. Temporal observational analysis showed that negative emoji usage during late-night and early-morning periods amongst depressed users corresponded with circadian rhythm disturbances characteristic of depression. Co-occurrence analysis demonstrated that emoji combinations yielded richer emotional context than individual emoji examination. We observed, based on frequency counts of emoji combinations within single posts, that depressed users' tendency to employ multiple

negative emojis (😔 + 😢) indicated emotional intensity levels beyond singular emoji capability. Context sensitivity analysis revealed emoji meaning heavily depended on surrounding textual content, validating the integrated analytical approach, rather than treating emojis as independent features. These patterns provide valuable insights into the relationship between digital expression modalities and underlying psychological states.

This investigation, as shown in Table 6, demonstrates the critical importance of MCC metrics for evaluating models on imbalanced datasets, rather than relying solely on accuracy measures. Furthermore, the integration of multi-modal frameworks [33] represents a significant avenue for future research development, which could substantially enhance models' capacity to identify individuals experiencing depression more accurately.

Table 6. Confusion matrix.

	Predicted Negative (Control)	**Predicted Positive (Depressed)**
Actual Negative (Control)	21,586 (TN)	1136 (FP)
Actual Positive (Depressed)	154 (FN)	2175 (TP)

5. Conclusions

This study demonstrates that the systematic integration of emoji features through descriptive text conversion substantially improves the performance of depression detection on text-based social media data. Our BERT–CNN–BiLSTM architecture with attention mechanisms achieved 97.12% accuracy, 94.56% precision, 93.44% F1-score, 85.67% MCC, and 91.23% AUC-ROC on the SuicidEmoji dataset, outperforming established baseline models. The emoji-to-text conversion strategy yielded +1.89% accuracy improvement over emoji removal, establishing that emoji preservation is beneficial, rather than supplementary, for mental health detection systems. Analysis revealed distinct emoji usage patterns: depressed users favour negative expressions (😔 13.9%, 😢 12.8%, 💔 6.7%), whilst controls prefer positive emojis (😂 16.5%, 😊 11.0%, 😎 10.2%). Attention mechanism visualisation demonstrated that the model learns to focus on clinically relevant features, including emotional indicators, personal pronouns, and emoji–text descriptors, providing interpretable insights into classification decisions.

Limitations: Several important limitations constrain the generalisability of our findings. Firstly, validation was conducted exclusively on Reddit data, limiting claims about cross-platform applicability. Reddit's pseudonymous environment, longer text formats, and specific community norms may not be generalisable to to Twitter (character limits), Instagram (image-centric), Facebook (real-name policies), or non-Western platforms (WeChat, Weibo). Secondly, the dataset exhibits demographic biases towards younger, male and Western English-speaking users, limiting its applicability to broader populations, including older adults, women, and non-English speakers. Thirdly, self-selection bias exists as r/SuicideWatch represents support-seeking individuals rather than the full spectrum of depression presentations, potentially excluding isolated or withdrawn individuals. Fourthly, classification relies on subreddit members rather than clinical diagnoses, introducing potential inaccuracies in self-labelling. Finally, we have not conducted formal ablation studies to quantify the contributions of individual architectural components, nor have we validated our results against contemporary transformer architectures (RoBERTa, DeBERTa, MentalBERT), which could limit our ability to definitively claim state-of-the-art performance beyond classical baselines.

Future Research Directions: Six priorities emerge for future work: (i) cross-platform validation across Twitter, Instagram, Facebook, and non-Western platforms to assess generalisability beyond Reddit's ecosystem; (ii) formal ablation studies systematically removing BERT, CNN, BiLSTM, and Attention components to quantify individual contributions and validate architectural necessity; (iii) benchmarking against advanced models, including RoBERTa, DeBERTa, domain-specific pre-trained models (MentalBERT), and recent emoji-aware architectures to position performance relative to contemporary approaches; (iv) multilingual evaluation with culturally diverse emoji datasets to address current Western English-speaking bias; (v) clinical validation with diagnosed populations and expert annotations to assess alignment with clinical depression criteria rather than self-reported subreddit membership; (vi) multimodal frameworks incorporating images, videos, and metadata alongside text and emojis for comprehensive mental health assessment. These directions would address current limitations and advance emoji-based depression detection toward robust, generalisable clinical applications.

Contribution: Despite these limitations, this work makes valuable contributions: demonstrating that emoji-to-text conversion outperforms removal in depression detection; providing the first systematic documentation of emoji distribution patterns in depression-related social media text data content; establishing attention-based interpretability for emoji feature contribution and achieving robust performance on imbalanced data through appropriate metric selection (MCC). These findings provide methodological guidance for researchers developing emoji-aware mental health detection systems, which we believe has strong future potential. This study also establishes proof-of-concept for treating visual emotional indicators as semantically equivalent to linguistic content within computational analysis frameworks.

Author Contributions: Conceptualization, J.P.T.; Methodology, J.P.T.; Validation, J.P.T.; Formal analysis, J.P.T.; Investigation, J.P.T.; Resources, J.P.T.; Data curation, J.P.T.; Writing—original draft, J.P.T.; Writing—review & editing, S.Y.; Visualization, J.P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original dataset presented in the study are openly available in https://github.com/gohjiayi/suicidal-text-detection at [27].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dattani, S.; Rodés-Guirao, L.; Ritchie, H.; Roser, M. Mental Health. Available online: https://ourworldindata.org/mental-health (accessed on 10 August 2025).
2. National Collaborating Centre for Mental Health. *Common Mental Health Disorders: Identification and Pathways to Care*; British Psychological Society: Leicester, UK, 2011.
3. Sinnema, H.; Terluin, B.; Volker, D.; Wensing, M.; van Balkom, A. Factors contributing to the recognition of anxiety and depression in general practice. *BMC Fam. Pract.* **2018**, *19*, 99. [CrossRef]
4. Mitchell, A.J.; Rao, S.; Vaze, A. International comparison of clinicians' ability to identify depression in primary care: Meta-analysis and meta-regression of predictors. *Br. J. Gen. Pract.* **2011**, *61*, e72–e80. [CrossRef] [PubMed]
5. De Choudhury, M.; Gamon, M.; Counts, S.; Horvitz, E. Predicting depression via social media. In Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, MA, USA, 8–11 July 2013; pp. 128–137.
6. Guntuku, S.C.; Yaden, D.B.; Kern, M.L.; Ungar, L.H.; Eichstaedt, J.C. Detecting depression and mental illness on social media: An integrative review. *Curr. Opin. Behav. Sci.* **2017**, *18*, 43–49. [CrossRef]

7. Ji, S.; Yu, C.P.; Fung, S.-f.; Pan, S.; Long, G. Supervised learning for suicidal ideation detection in online user content. *Complexity* **2018**, *2018*, 6157249. [CrossRef]
8. Tadesse, M.M.; Lin, H.; Xu, B.; Yang, L. Detection of depression-related posts in reddit social media forum. *IEEE Access* **2019**, *7*, 44883–44893. [CrossRef]
9. Du, Y. The impact of emojis on verbal irony comprehension in computer-mediated communication: A cross-cultural study. *Int. J. Hum. Comput. Interact.* **2025**, *41*, 4979–4986. [CrossRef]
10. Miller, H.; Thebault-Spieker, J.; Chang, S.; Johnson, I.; Terveen, L.; Hecht, B. "Blissfully happy" or "ready tofight": Varying interpretations of emoji. In Proceedings of the International AAAI Conference on Web and Social Media, Cologne, Germany, 17–20 May 2016; pp. 259–268.
11. Wiederhold, B.K. Modern Hieroglyphics and the Generation Gap: Do Emojis Need Their Own Rosetta Stone? *Cyberpsychology Behav. Soc. Netw.* **2024**, *27*, 167–168. [CrossRef] [PubMed]
12. Balan, A.; Tahir, R. Visual Cues in Survey Design-A Strategic Use of Emojis in Research. In Proceedings of the International Conference on Human-Computer Interaction, Gothenburg, Sweden, 22–27 June 2025; pp. 13–30.
13. Sia, J.K.-M.; Hii, I.S.; Jong, L.; Low, W.W. Do emojis really help us to communicate better? Investigating instructor credibility, students' learning motivation, and performance. *Educ. Inf. Technol.* **2024**, *29*, 17889–17913. [CrossRef]
14. Marengo, D.; Settanni, M.; Giannotta, F. Development and preliminary validation of an image-based instrument to assess depressive symptoms. *Psychiatry Res.* **2019**, *279*, 180–185. [CrossRef]
15. Thekkekara, J.P.; Yongchareon, S.; Liesaputra, V. An attention-based CNN-BiLSTM model for depression detection on social media text. *Expert Syst. Appl.* **2024**, *249*, 123834. [CrossRef]
16. Carroll, J. The role of prosocial behaviour, personality and general mental health in predicting emoji use and preference. *Psychol. Rep.* **2025**, *128*, 4210–4226. [CrossRef] [PubMed]
17. Settanni, M.; Marengo, D. Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Front. Psychol.* **2015**, *6*, 1045. [CrossRef] [PubMed]
18. Kralj Novak, P.; Smailović, J.; Sluban, B.; Mozetič, I. Sentiment of emojis. *PLoS ONE* **2015**, *10*, e0144296. [CrossRef]
19. Bhargava, N.; Radaideh, M.I.; Kwon, O.H.; Verma, A.; Radaideh, M.I. On the Impact of Language Nuances on Sentiment Analysis with Large Language Models: Paraphrasing, Sarcasm, and Emojis. *arXiv* **2025**, arXiv:2504.05603. [CrossRef]
20. Jain, N.; Wu, Z.; Villalobos, C.E.M.; Hilliard, A.; Guan, X.; Koshiyama, A.; Kazim, E.; Treleaven, P.C. *From Text to Emoji: How PEFT-Driven Personality Manipulation Unleashes the Emoji Potential in LLMs*; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 4687–4723.
21. Cui, S.; Thakur, N.; Poon, A. Emoji Retrieval from Gibberish or Garbled Social Media Text: A Novel Methodology and a Case Study. In Proceedings of the International Conference on Human-Computer Interaction, Washington, DC, USA, 29 June 2024; pp. 170–189.
22. Ali, M.M.; Mohamed, E.S. DGDO-BiLSTM: Dominance guiding defense optimization-based bidirectional long short-term memory for sentiment analysis using multilingual text and emojis. *Inf. Sci.* **2025**, *716*, 122193. [CrossRef]
23. Barbieri, F.; Kruszewski, G.; Ronzano, F.; Saggion, H. How cosmopolitan are emojis? Exploring emojis usage and meaning over different languages with distributional semantics. In Proceedings of the 24th ACM international Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 531–535.
24. Saeidi, S. Identifying personality traits of WhatsApp users based on frequently used emojis using deep learning. *Multimed. Tools Appl.* **2024**, *83*, 13873–13886. [CrossRef]
25. Gollapalli, S.D.; Ng, S.-K. Modeling Emoji Generation for Emotion Analysis of Social Media Short Texts. In Proceedings of the ICWSM Workshops, Online, 7–10 June 2021.
26. Li, J. Sentiment analysis of text based on emoji attention mechanisms: A new approach to online course evaluation. *Int. J. Inf. Commun. Technol.* **2025**, *26*, 70–86. [CrossRef]
27. Zhang, T.; Yang, K.; Ji, S.; Liu, B.; Xie, Q.; Ananiadou, S. SuicidEmoji: Derived Emoji Dataset and Tasks for Suicide-Related Social Content. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024; pp. 1136–1141.
28. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
29. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *28*, 2222–2232. [CrossRef]
30. Zulqarnain, M.; Ghazali, R.; Hassim, Y.M.M.; Rehan, M. A comparative review on deep learning models for text classification. *Indones. J. Electr. Eng. Comput. Sci* **2020**, *19*, 325–335. [CrossRef]
31. Huang, J.; Ling, C.X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310. [CrossRef]

32. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
33. Gui, T.; Zhu, L.; Zhang, Q.; Peng, M.; Zhou, X.; Ding, K.; Chen, Z. Cooperative multimodal approach to depression detection in twitter. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 110–117.