An Approach for Emotion Detection Using a Hybrid Vision Sequence-Based Convolutional Neural Networks (Vi-CNN) Model

Vian Sabeeh
Information technology management
Technical College of Management
Middle Technical University
Baghdad, Iraq
viantalal@mtu.edu.iq
Orcid.org/0000-0002-0860-2335

Ahmed Bahaaulddin A Alwahhab
Information technology management
Technical College of Management
Middle Technical University
Baghdad, Iraq
ahmedbahaaulddin@mtu.edu.iq
Orcid.org/0000-0003-0965-4812

DOI: http://dx.doi.org/10.31642/JoKMC/2018/120205
Received Feb.21, 2025. Accepted for publication Jun. 24, 2025

***Abstract*—*Emotion represents a significant reflection or indicator of human mental states, biological effects, and physiological states, and it also plays a vital role in human interpersonal communication and decision-making. People convey their emotions naturally via everyday interactive communication due to the increasing advancement of social media platforms. Nowadays, detecting emotions from extensive textual data helps to provide expressive information for understanding the behavior of humans. Meanwhile, most prior techniques used for emotion detection are insufficient in providing promising results from long-term contextual information. Thus, it motivates introducing a hybrid deep learning model, Vision sequence-based Convolutional Neural Network (Vi-CNN), from text data to detect emotion. The Bidirectional Encoder Representation from Transformer (BERT) transforms the input texts into tokens. Then, extracting the most suitable and appropriate features from the text tokens is executed. Following this, the Vi-CNN model significantly detects emotions from the extracted features and classifies the recognized emotions. Furthermore, the results obtained by Vi-CNN are compared with those of other prevailing schemes. The experimental results highlight that the Vi-CNN attained promising results with a maximum recall of 94.765%, a precision of 92.988%, and an F-measure of 93.867%.*

Keywords—Deep learning; emotion recognition; Vision transformer; Convolutional Neural Network; text data.

I. INTRODUCTION

Emotions are a significant aspect of human life and play a major role in how individuals understand and perceive their surroundings [1]. Emotion is a dynamic physiological and cognitive state that arises in response to stimuli, such as interactions, thoughts, or experiences with others. It encompasses communication, physiological responses, behavioral influences, cognitive processes, and subjective experiences [2]. Emotions are triggered by changes occurring in various biological systems within the human body. Automatic emotion recognition has enabled numerous applications, including marketing analysis based on emotional responses, post-traumatic rehabilitation, and psychological treatment [3]. Emotions like happiness, anger, fear, and sadness are frequently experienced daily [4]. In recent years, feelings, opinions, and emotions have been shared by users on social media platforms like YouTube, Twitter, and Facebook due to the rapid growth of multimodal social media applications [5]. Moreover, security, politics, healthcare, psychology, and business organizations have gathered people's emotions through social interactions [4]. Recognizing emotions is essential due to various feelings, including stress. Health psychologists study emotion extraction to assist patients by establishing connections between emotions, anxiety, and health [6][7].

Speakers' feelings are recorded in various media formats through social media, including videos, audio, and text, to recognize emotions accurately [8]. These formats effectively reflect a person's emotional state. Text is often considered the most effective medium for conveying emotions [9][10]. Typically, individuals share their feelings through blogs, comments, status updates, and posts on social media platforms. To identify the user's genuine emotion, the sentiment conveyed in their posts is analyzed thoroughly [7]. Emotion recognition from text-based conversations has garnered increased attention due to its potential applications in opinion mining and sentiment analysis of publicly available conversational data [11]. In recent years, a wealth of textual data has emerged online, making it enjoyable to extract emotions from this data for various purposes, including business [12]. Recognizing emotions in text involves assessing the feelings expressed to comprehensively understand their intention, expression, and sentiment. The primary mechanism for detecting emotional content in text data relies on Natural Language Processing (NLP). NLP is a developing field that has gained traction with

the increasing volume of online comments [13]. The recognition of emotions in text plays a crucial role in human-computer interaction tasks. Several steps, such as preprocessing, feature extraction, ranking, classification, and validation, are employed in NLP to detect emotions more effectively [14]. However, emotion detection faces challenges when multiple emotions are present within a single comment [7].

Emotion recognition has gained significant attention among researchers over the past decades in textual conversations to identify individual opinions, emotions, and sentiments by considering specific issues and policies. The user experience, communication, and personalization are enhanced by accurately detecting and understanding emotional content in text. The baseline approaches used for emotion detection were crucial in identifying emotions from imbalanced databases, which also achieved notable performance in capturing complex emotions from textual dialogues. However, these approaches faced various challenges that affected the accuracy and effectiveness of detected emotions. Consequently, this motivates the development of a hybrid deep-learning scheme for emotion detection. Researchers propose hybrid models, including learning-based, lexical affinity, and keyword-based techniques to identify textual emotions [15]. Moreover, researchers have combined recognition techniques to design hybrid models that have achieved highly accurate results in recent years [16]. Additionally, these approaches encountered various complexities in recognizing textual emotions and addressed numerous issues [17]. Some significant challenges in recognizing emotions from abbreviated and short texts include error analysis, which is also insufficient for effective feature extraction of special symbols and emojis from text [7]. Recent advancements in Artificial Intelligence (AI) techniques have utilized machine learning and deep learning methods to classify emotions in various formats [16]. Different emotion classification techniques employ machine learning to analyze underlying patterns in unlabeled training data. Deep learning methods have achieved superior accuracy compared to other machine learning techniques while recognizing texts from large datasets [16]. The recent growth of deep learning models has supported many valuable applications across various domains and yielded promising results for decision-making, voice recognition, object detection, and pattern classification [18]. Deep learning techniques, such as Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Deep Boltzmann Machine (DBM), and Deep Belief Networks (DBN), are commonly used for emotion recognition and have demonstrated superior performance in providing solutions to various NLP tasks, such as text classification, machine translation, and language modeling [4][19].

This paper proposes Vi-CNN for detecting emotions from input text data. The acquired text data is initially sent to BERT tokenization to transform into tokens. Later, various features are extracted from these text tokens using different text feature extractors. Following this, emotion detection is performed using Vi-CNN on the extracted features, and the recognized emotion is finally classified into various categories.

The contribution of this research is summarized as,

- Introduced Vi-CNN for emotion detection. It is developed by incorporating deep learning techniques like Vision Transformer (ViT) and sequence-based Convolutional Neural Networks (CNN) to automatically detect emotion from text data.

The remaining portion of the manuscript is organized as follows: Section II presents the relevant work on baseline approaches, and Section III explains the proposed Vi-CNN approach developed for emotion detection. Additionally, Section IV illustrates the outcomes and discussions that ensued, while Section V summarizes the manuscript's conclusion and discusses future research.

II. RELATED WORK

Tu, G., et al. [20] designed a Context- and Sentiment-Aware Graph Attention (CSAGAT) technique to recognize emotion from textual conversation. This research used a dialogue transformer based on hierarchical multi-head attention to determine the inter- and intra-dependency relationships from contextual utterances. The sentiment- and context-aware graph attention mechanism was also used to represent commonsense knowledge dynamically. This approach effectively enhanced the sentimental intensity and modeled the information for emotion recognition. However, this approach failed to capture relationship dependency among the speakers, so a graph convolutional network (GCN) was utilized to increase the performance further. Mohammad, F., et al. [11] introduced a Text Augmentation-based computational model (TA-MERT) to recognize emotions via transformers. Here, the transformer encoder captured the backward and forward contextual information, specifically the transformer-based BERT model. This approach significantly handled the relationships and patterns in the data, but it failed to prevent data overfitting issues and enhance the accuracy of the validation. Asghar, M.Z., et al. [21] developed a Bidirectional Long-Term Short-Term Memory (BiLSTM) model to recognize emotions like guilt, shame, fear, sadness, and joy. This model utilized backward and forward LSTM to learn contextual information from backward and forward directions. This model significantly reduced the misclassification rate by capturing the semantics of words. Meanwhile, this approach failed to increase the performance by considering different emotion intensities, like weak positive, strong positive, weak negative, and strong negative. Han, T., et al. [13] established the XLNet Bidirectional Gated Recurrent Unit and Attention (XLNet-BiGRU-Att) model for improving the performance of text emotion recognition. Here, the contextual information was learned for building bidirectional language models using XLNet, and more compelling features were extracted using a Bidirectional Gated Recurrent Unit (BiGRU). This model significantly reduced the complexities and processed long sequence data for emotion recognition, but this approach did not successfully prevent memory overflow issues during emotion detection.

Wen, J., et al. [8] introduced the Dynamic Interactive Multiview Memory Network (DIMMN) to recognize emotions by incorporating interaction information. In this approach, the

emotional impacts of the speakers were collected using a Gated Recurrent Unit (GRU) and Temporal Convolutional Network (TCN) to select various global information based on queries. This method increased the robustness and accuracy during emotion detection by solving information fusion issues. Meanwhile, this model failed to utilize more modalities to perform interactive Multiview learning to improve emotion recognition further. Shelke, N., et al. [7] designed a Leaky Relu Activated Deep Neural Network (LRA-DNN) to analyze emotion effectively. This model significantly reduced the misclassification and misprediction errors while detecting emotions. However, this model failed to consider advanced deep learning algorithms for increasing the recognition performance. Bharti, S.K., et al. [16] devised CNN+Bi-GRU+SVM for text-based emotion recognition. This method effectively detected emotions from multi-text sentences, tweets, dialogs, keywords, and lexicon words, but this approach was unsuitable for real-world applications to recognize texts. Ghafoor, Y. et al. [22] introduced Textual Emotion Recognition in Multidimensional Space (TERMS) for handling emotional boundaries. Textual Emotion Recognition in Multidimensional Space (TERMS) was used in this model to handle texts' semantic meaning and syntactic structures for exposing linguistic and contextual information. This method was more adaptable and achieved high precision values during the detection of emotions, but this technique failed to increase recognition performance by utilizing deep learning models.

The drawbacks faced by prevailing techniques during the detection of emotion from texts are listed as follows,

- The CSAGAT model in [20] effectively incorporated sentiment and contextual awareness information to recognize emotion in dialogues, increasing the model's responsiveness to context and sentiment. Meanwhile, this approach failed to consider multimodal features for emotion recognition that involve integrating data from different modalities and sources to increase recognition accuracy.
- The TA-MERT technique used in [11] significantly recognized human emotions during textual conversation, but this model failed to increase emotion recognition reliability significantly.
- The BiLSTM approach utilized in [21] enhanced the understanding of context by capturing information from past and future sequences. However, it failed to consider the NLP features that increase the robustness and accuracy of emotion recognition systems.
- The DIMMN model used in [8] effectively leveraged the multiple conversation modalities to increase the emotion recognition performance. Meanwhile, this technique failed to increase emotion recognition accuracy by enhancing the richness of input data.
- The existing techniques used for emotion recognition did not successfully capture the personal differences among emotions to completely cover the possible emotional contents embedded in a text. These techniques also encountered difficulties aligning features from various modalities for accurate emotion recognition.

III. METHODOLOGY

This paper proposes Vi-CNN for detecting emotions from text data. At first, the text data is acquired from the emotion detection text dataset [23], and the acquired data is tokenized using the BERT tokenization technique [24]. Later, different features, namely Term Frequency-Inverse Document Frequency (TF-IDF) features [25], length of the text, punctuation marks, hashtags, capitalized words, a bag of words, SentiWordNet features [26], and similarity score [27], are extracted from the resultant tokenized texts. Following this, the detection of emotion is performed using Vi-CNN from the extracted features, and the detected emotion is classified as surprise, sadness, relief, anger, neutral, love, boredom, enthusiasm, fun, happiness, hate, and worry. Here, the Vi-CNN is developed by integrating a sequence-based CNN [28] and ViT [29]. Fig.1 shows the schematic view of the Vi-CNN for emotion detection.

A. Input text data acquisition

The input text for the emotion detection task is collected from annotated tweets with various emotions, and the database comprises 13 types of emotions, with 40000 records under three columns, like content, sentiment, and Tweet_ID. The “sentiment” consists of various emotions behind the text, whereas the “content” consists of raw tweets. The dataset is given as in (1),

$$T = [T_1, T_2, T_3, .., T_D, \ldots, T_W] \quad (1)$$

Where, T is the text dataset considered for emotion detection, W signifies the number of texts available in the Emotion detection from text dataset, and T_D represents the D^{th} text accumulated for the detection task.

B. BERT tokenization

Once the input text is selected for the emotion detection

task, the selected text is converted into a suitable form, like

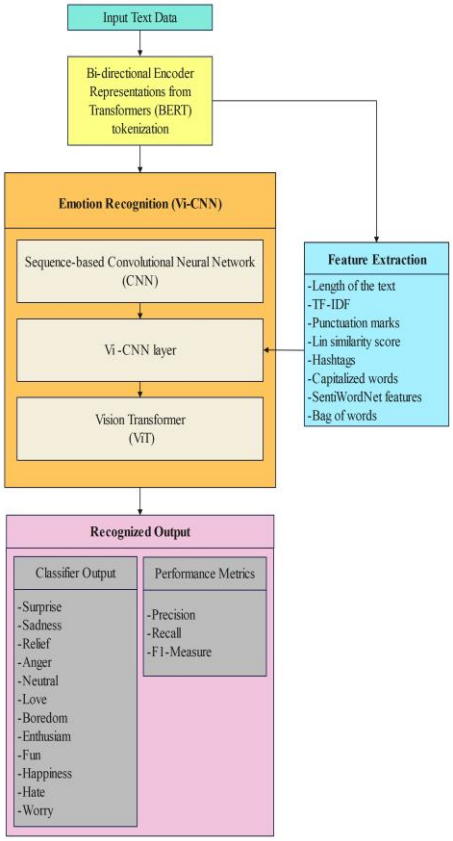


Fig. 1. Schematic view of the Vi-CNN for emotion detection from text data

tokens that machines can easily understand and process. Here, the selected input text T_D is subjected to BERT tokenization [24] for converting input text into tokens through bidirectional self-attention. The BERT tokenization method quickly transforms extensive sentences into tokens, which is also more effective in handling short input sequences. In BERT, the transformer encoder captures text contextual information through a self-attention mechanism. Then, it generates a contextual embedding vector sequentially. Moreover, the BERT performs various downstream tasks to represent single and paired sentences, where the paired sentences are packed into a single sentence. After the representation of sentences, the first token of all sequences is used to select the special classification token, and the aggregate sequence representation is executed based on the final hidden state next to the first token. Following this, the tokens are generated by separating

the sentence and then embedded to get text sequences in a tokenized form. Hence, the resultant token B_T is attained during BERT tokenization from the selected input text T_D.

C. Feature extraction

After the generation of tokens, the different features are extracted from the resultant tokens B_T different extractors transfer the raw data into a suitable form to perform accurate detection tasks. In this research, the features, like TF-IDF features [25], length of the text, Lin similarity score [27], and SentiWordNet features [26], like punctuation marks, hashtags, capitalized words, and a bag of words, are extracted from the resultant token B_T using different extractors. The extraction process performed by each feature extractor is delineated below:

1) Length of the text: The length of the text [30] is defined as the number of words available in the sentence, which is measured by taking the ratio of total words available in the sentence to the total words available in longest sentence of the document. The length of the text is designed in (2) as follows,

$$M_1 = \frac{Total\ words\ available\ in\ sentence}{Total\ available\ in\ the\ longest\ sentence} \quad (2)$$

here, M_1 denotes the extracted length of the text.

2) TF-IDF: The TF-IDF feature [25] reads the document effectively through text summarization. Generally, the TF-IDF is a weighting parameter used to assign weights to terms. The TF refers to the frequency of terms appearing in a document and the number of words presented. The TF is computed by dividing total data instances using data count and is given as in (3),

$$TF, \alpha = \frac{U}{V} \quad (3)$$

where the number of times terms appear in a document is given by V, and the number of words in a document is represented as U. Further, the IDF is expressed as in (4),

$$IDF, \beta = log\frac{I}{J} \quad (4)$$

here, the total utilized documents are signified as I and the total documents with the presence of the selected term are given by J. Hence, the extracted TF-IDF feature from the token B_T is expressed by (5),

$$M_2 = \alpha \times \beta \quad (5)$$

where, M_2 symbolizes the TF-IDF extracted feature.

3) Lin similarity: The Lin similarity [27] is used to identify the semantic similarity between two words by considering the commonness between the words and the described information. The Lin similarity is given as in (6),

$$M_3 = 2 * \frac{\chi\left(\delta(H_1, H_2)\right)}{\chi(H_1) + \chi(H_2)} \quad (6)$$

here, the extracted Lin similarity feature is represented as M_3, the lowest common subsumer is given byδ, the information content is signified asχ, $H_1$and H_2 symbolizes word 1 and word 2.

4) SentiWordNet features: The SentiWordNet [31] is considered a lexical resource used for opinion mining, where the SentiWordNet uses three sentiment numerical scores, such as Neg(s), Pos(s), and Obj(s) to each synsets of WordNet to

describe how negative, positive, and objective terms are available in synsets. The extracted SentiWordNet feature from the token B_T is signified as M_4.

5) Punctuation marks: The exclamation marks, apostrophes, dots, etc, available in the document are termed punctuation marks [26] and are expressed as in (7),

$$M_5 = \sum_{k=1}^{w} A_k^g \quad (7)$$

where, M_5 indicates the extracted punctuation feature, A_k^g indicates the total punctuations presented in the g^{th} review, and w symbolizes the total words presented in the text.

6) Hashtag: generally, hashtags [26] are utilized on Twitter to get specific information by contributing new things, and the extracted hashtag feature is given by (8) ,

$$M_6 = \sum_{k=1}^{w} C_k^g \quad (8)$$

where, M_6 indicates the extracted Hashtags feature, and C_k^g represents the hashtags presented in g^{th} review.

7) All-caps: all-caps [26] is the total capitalized words presented in the text document and is expressed as in (9),

$$M_7 = \sum_{k=1}^{w} D_k^g \quad (9)$$

here,M_7signifies the extracted all-cap feature and the total capitalized words presented in g^{th} review is represented as D_k^g.

8) Bag of word: In Bag of Word [32], the documents are considered a collection of words that help to represent the text data regardless of order or grammar. The Bag of Word handles each document as a numerical vector set with a fixed length, and each feature for the extraction of the Bag of Word feature represents the frequency of occurrence of each word M_8.

Thus, the final extracted text feature obtained from all the features extracted from the token B_T are given as in (10),

$$M_{Extract} = [M_1, M_2, M_3, \ldots, M_8] \quad (10)$$

Later, the final extracted features $M_{Extract}$ from the token B_T Emotion recognition tasks using Vi-CNN are allowed.

D. Emotion detection using Vi-CNN

Emotion is considered a significant indicator of mental states and biological effects and also explains the physiological statements of humans. Emotion recognition plays a vital role in accurately detecting and understanding text emotions. In this research, emotion detection is performed using the Vi-CNN model, where the Vi-CNN is developed by integrating deep learning approaches, like Sequence-based CNN [28] and ViT [29]. The emotion recognition task in Vi-CNN is carried out under the Sequence-based CNN model, the Vi-CNN layer, and the VisionNet model. Moreover, harmonic analysis [33] is used in the fusion and regression modeling layer in Vi-CNN to enhance its strength. The input data T_D is fed into a Sequence-based CNN model to get the output γ_1. After that, the output of the Sequence-based CNN model γ_1 and the extracted feature $M_{Extract}$ are fed into the Vi-CNN layer to obtain the output γ_2. The resultant output is fed into ViT to identify the final detected emotion γ_3. Moreover, Fig. 2 shows the systematic view of the Vi-CNN approach for emotion detection.

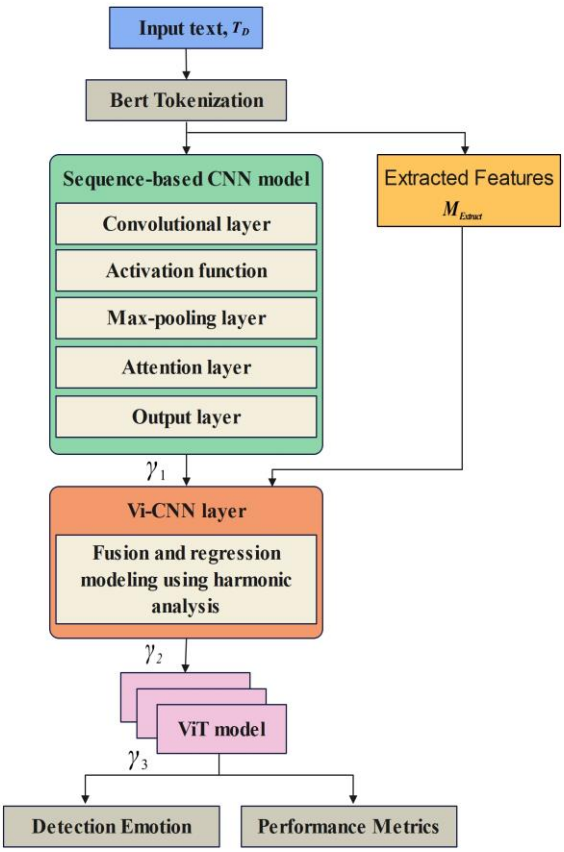


Fig. 2. Block diagram of Vi-CNN approach for emotion detection

1) Sequence-based CNN

In sequence-based CNN [28], the corpus is trained using the CNN model with an attention mechanism. The corpus mainly comprises a sequence of utterances that helps to provide solutions for the sequence of detection tasks. The sequence-based CNN utilizes input text data from the present sentence for emotion detection and multiple feature extraction channels for emotion detection. This model comprises a one-dimensional convolutional layer with a non-linear activation function, which utilizes max pooling operations to extract features from a fully connected layer with a SoftMax classifier and input data for accurate classification of emotion classes. Moreover, the filters are convolved over the word sequences to extract features. Also, the convolutional layer extracts the

context and local features using vector representation, and the extracted features are passed via the max pooling layer to the fully connected layer. Later, the context CNN features are sent to the attention model output to capture and classify the most important semantic information from the sentence. The sequence-based CNN mainly comprises five layers, such as the convolutional layer, fully connected layer, pooling layer, and attention mechanism layer, where the process carried out in each layer is explicated below.

- Convolutional layer: is the core sequence-based CNN, which helps to perform convolution filter operations to extract features from input text data T_D. It employes filters sweeping across all sentences, convolving with words to produce feature maps. The weight of the neurons is updated through the training session, while the filters are randomly initialized.
- Activation function: Among all activation functions, such as Hyperbolic Tangent (tanh), Rectified Linear Unit (ReLU), and sigmoid function, the ReLU is the most commonly used activation function for increasing the training speed and accuracy of the results. In a sequence-based CNN, the nonlinear data, such as a single line, cannot be easily separated once a nonlinear activation function handles classification. The ReLU activation function is given as in (11),

$$f(T_D) = Max(0, T_D) \tag{11}$$

- Max pooling: The activation function output is subjected to the pooling layer once the ReLU activation function is performed. A max pooling operation is applied to the outcome of the convolutional layer, and the resultant matrix of the pooled feature map is expressed as in (12),

$$W = \left[a(\varepsilon(\phi_1)), a(\varepsilon(\phi_2)), a(\varepsilon(\phi_3)), \ldots, a(\varepsilon(\phi_N))\right] \tag{12}$$

  Here, the feature map is signified as ϕ_N, the ReLU operation is represented as ε, and the pooling operation is given by a. Then, the output of the CNN layer is reshaped by subjecting the feature map to a fully connected layer.

- Attention model: The output vectors $(z_1, z_2, z_3, \ldots z_l)$ generated by the fully connected layer are taken as the matrix, where the sentence length is represented l. Later, a vector is returned, and the output vector summary focuses on information linked to the context. Then, the weighted arithmetic mean of the output vectors is returned. Moreover, the weights are selected based on the relevance of each output vector according to the given context. Some of the steps taken under consideration are, at first, the feature vectors and the input context are accurately recognized, and then, the output vector and context aggregation are performed. After that, a SoftMax is used to compute the weights and the maximum relevance of the variables based on the contexts. Finally, the output is taken as the weighted arithmetic mean of all output vectors, and the most correlated variables are selected by the attention model with the context.
- Output layer: The attention layer output is fed to the output layer, which generates N-dimensional vectors. Later, a SoftMax classifier is utilized to determine the probabilities for predicting output emotions. The resultant output obtained from the SoftMax layer is expressed as in (13),

$$\gamma_1 = softm\,ax(T_D * L_c + \psi_c) \tag{13}$$

where the bias is symbolized as ψ_c, the weight vector is indicated as L_c, and the output of Sequence-based CNN is given by γ_1. Further, the architecture of Sequence-based CNN is displayed in Fig. 3.

2) Vi-CNN layer

The output of the Sequence-based CNN model $\gamma_1$and extracted feature $M_{Extract}$ fed and fused using harmonic analysis [33] In the Vi-CNN layer. Here, the unknown deterministic elements are determined using harmonic analysis from the models with unknown amplitudes and frequencies. It enables the combination of the two inputs into a unified output, thereby enhancing recognition accuracy. The time series is designed based on periodical structures. At first, the Vi-CNN layer output is obtained based on the extracted feature $M_{Extract}$ at q^{th} time interval, which is expressed as in (14) ,

$$y_1 = \sum_{c=1}^{n} (M_{Extract})_c\, \vartheta_c \tag{14}$$

here, the extracted features are represented as $M_{Extract}$, and the weight coefficient is indicated by ϑ. Later, the time series is modeled to perform harmonic analysis based on the periodical structures, which is given by (15),

$$E_{(1)}, E_{(2)}, E_{(3)}, \ldots, E_{(i)}, \ldots, E_{(w)} \tag{15}$$

where, the i^{th}time series data taken for observation is represented as $E_{(i)}$, and the length of the time series is given by w.

After that, the description of the observed series $E_{(i)}$ is obtained by an orthogonal trigonometric function. Hence, if E is odd, then $w = 2b + 1$and the orthogonal trigonometric function is expressed as (16),

$$E_{(i)} = \varpi_0 + \sum_{j}^{b} \left(\varpi_j\, cos\left(\frac{2\pi ji}{w}\right) + \lambda_j\, sin\left(\frac{2\pi ji}{w}\right)\right) \tag{16}$$

where the total cycles are represented as b, ϖ and λ are the preselected constants [34].

Let us consider, $w = 2$ and $b = 1$, the resultant equation (16) is given by (17) and (18),

$$E_{(i)} = \varpi_0 + \varpi_1\, cos\left(\frac{2\pi i}{2}\right) + \lambda_1\, sin\left(\frac{2\pi i}{2}\right) \tag{17}$$

$$E_{(i)} = \varpi_0 + \varpi_1 \cos \pi\, i + \lambda_1 \sin \pi\, i \tag{18}$$

here,

$$\varpi_0 = \frac{1}{w}\sum_{i=1}^{w} E_{(i)} \quad (19)$$

$$\varpi_0 = \frac{1}{2}\left[E_{(1)} + E_{(2)}\right] \quad (20)$$

$$\varpi_j = \frac{2}{w}\sum_{i=1}^{w} E_{(i)}\, cos(2\pi ji/w) \quad (21)$$

Assume, $j = 1, = 2$, the equation (22) is written as,

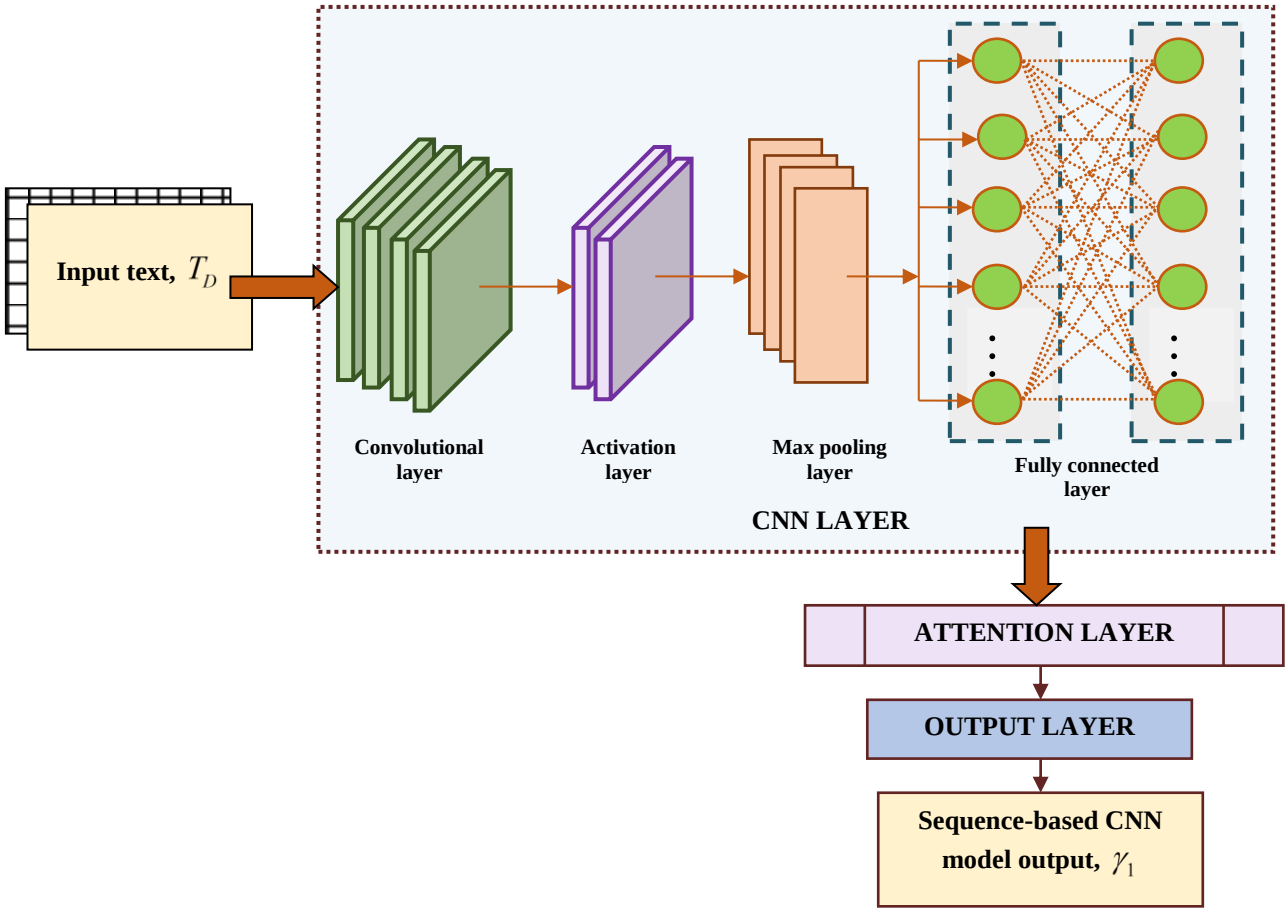


Fig. 3. Architecture of the Sequence-based CNN model

$$\varpi_1 = \frac{2}{2}\left[E_{(1)}\, cos\left(\frac{2\pi}{2}\right) + E_{(2)}\, cos\left(\frac{2\pi 2}{2}\right)\right] \quad (22)$$

$$\varpi_1 = E_{(1)}\, cos(\pi) + E_{(2)}\, cos(2\pi) \quad (23)$$

By substituting the values of $cos(\pi)$and $cos(2\pi)$ in equation (23), and the equation is written as in (24),

$$\varpi_1 = E_{(1)}(-1) + E_{(2)}(1) \quad (24)$$

$$\varpi_1 = -E_{(1)} + E_{(2)} \quad (25)$$

Furthermore, the expression of λ_j is expressed by (26),

$$\lambda_j = \frac{2}{w}\sum_{i=1}^{w} E_{(i)}\, sin(2\pi ji/w) \quad (26)$$

As $j = 1, w = 2$,

$$\lambda_1 = \frac{2}{2}\left[E_{(1)}\, sin\left(\frac{2\pi}{2}\right) + E_{(2)}\, sin\left(\frac{2\pi 2}{2}\right)\right] \quad (27)$$

$$\lambda_1 = E_{(1)}\, sin(\pi) + E_{(2)}\, sin(2\pi) \quad (28)$$

Substituting the values of $sin(\pi)$ and $sin(2\pi)$ in equation (28), and the resultant equation is given as in (29),

$$\lambda_1 = E_{(1)}(0) + E_{(2)}(0) \quad (29)$$

where the time series model is signified as $E(i-1), E(i)$, and $E(i+1)$. Thus, $E_{(2)} = E(i), E_{(1)} = E(i-1)$, and $E_{(i)} = E(i+1)$.

Hence, by substituting the expression of $E_{(1)}, E_{(2)}$, and $E_{(i)}$ in (20), (25), (28), and (29), the resultant expression is given as in (30),

$$E(i+1) = \varpi_0 + \varpi_1\, cos\,\pi\, i + \lambda_1\, sin\,\pi\, i \quad (30)$$

$$\varpi_1 = -E(i-1) + E(i) \quad (31)$$

$$\lambda_1 = 0 \quad (32)$$

Further, after derivation process, we get:

$$E(i+1) = E(i)\left[\frac{1+2\,cos\,\pi i}{2}\right] + E(i-1)\left[\frac{1-2\,cos\,\pi i}{2}\right] \quad (33)$$

Consider, $E(i) = y_1, E(i-1) = \gamma_1$, and $E(i+1) = \gamma_2$, we get,

$$\gamma_2 = y_1\left[\frac{1+2\,cos\,\pi i}{2}\right] + \gamma_1\left[\frac{1-2\,cos\,\pi i}{2}\right] \quad (34)$$

$$\gamma_2 = \sum_{c=1}^{n}(M_{Extract})_c\, \vartheta_c\left[\frac{1+2\,cos\,\pi i}{2}\right] + (softmax(T_D * L_c + \psi_c))\left[\frac{1-2\,cos\,\pi i}{2}\right] \quad (35)$$

where the resultant output of the Vi-CNN layer is signified as γ_2, which is further fed into the ViT model to get the final detected emotion.

3) ViT model

Nowadays, the ViT model [29] is gaining massive attention in the vision community and is used to perform large vision application tasks. The ViT model performs intra- and inter-calculations between tokens by providing solutions to underlying tasks and splitting the output of the Vi-CNN layer γ_2 into ordered patches in series.

Assume a ViT model is applied with the Vi-CNN layer. γ_2 of height η, width κ, and channel ρ to execute the emotion detection task. Hence, the generated output is expressed as in (36),

$$\gamma_3 = \rho \circ X^h \circ X^{h-1} \circ \cdots \circ X^1 \circ \gamma_2 \quad (36)$$

here, $X(.)$ represents the transformer block at the layer h. Thus, by substituting γ_2 in (36), and the resultant equation is written as in (37),

$$\gamma_3 = \rho \circ X^h \circ X^{h-1} \circ \cdots \circ X^1 \circ \left(\sum_{c=1}^{n}(M_{Extract})_c\, \vartheta_c\left[\frac{1+2\,cos\,\pi\, i}{2}\right] + (softmax(T_D * L_c + \psi_c))\left[\frac{1-2\,cos\,\pi\, i}{2}\right]\right) \quad (37)$$

Let us consider that the transformer block is used to transform all the tokens from $(H-1)^{th}$ layer to H^{th} layer, and the process is expressed as in (38),

$$Y_{1:g}^{H} = X^H\left(Y_{1:g}^{H-1}\right) \quad (38)$$

where, $Y_{1:g}^{H}$ indicates the modified token, and the number of tokens is indicated as I. Later, the dimension of consistent feature G is used for all tokens by ViT throughout the layers, making the model jointly monitor all layers and making it easy to learn and capture global halting mechanisms. Therefore, the final detected emotion γ_3 is obtained using ViT, and Fig. 4 shows the structure of the ViT [35].

Further, the detected emotions are finally classified into different classes: anger, boredom, enthusiasm, emptiness, fun, sadness, neutral, hate, relief, love, happiness, surprise, and worry. Table I delineates the results obtained from the experiment by Vi-CNN on classifying emotions, such as anger, boredom, enthusiasm, emptiness, fun, sadness, neutral, hate, relief, love, happiness, surprise, and worry.

E. Practical Implementation

Due to the complex computations involved in the practical implementation of the proposed model, we explain each step in the model using a small simulation example.

TABLE I. CLASSIFIED EMOTION BY VI-CNN

Input text	Classified emotion
I had a dream about a pretty beach, and there was no beach when I woke up	Surprise
fuckin' trans telecom	Anger
Pats in Philly at 2 a.m.- I love it. Mmm, cheesesteak. I miss my boyfriend, but I love vacation.	Love
It is so annoying when she starts typing on her computer in the middle of the night!	Hate
bed...sorta. today was good; Sara has strep, though Angelina does too. I shared water with her B4, and they told me I will prob get it to	Enthusiasm
Waiting in line @ tryst	Boredom
Wondering why I'm awake at 7 am, writing a new song, plotting my evil secret plots muahahaha...oh damn it, not secret anymore	Fun
# I am excited to join you tomorrow	Happiness
I have a headache, so I'm going to bed. Goodnight!	Empty
Funeral ceremony...gloomy Friday...	Sadness
SoCal! stoked. or maybe not. tomorrow	Neutral
I'm at work	Relief
Choked on her retainers	Worry

Suppose the text input data is "# I am excited to join you tomorrow", and the emotion classification classes are [love, sadness, happiness, and neutral].

1. The model starts tokenizing and converting each token into a vector for the input text using BERT tokenization. Where the vocabulary size for each token's vector is 3. Table II shows the BERT tokenization operation.

2. In this step, the engineering features, such as capitalized words, hashtags, and text length, are extracted from the BERT tokenization. In this example, three features are extracted for simple illustration; Therefore, $M_{extracted} = [1, 1, 8]$. Where in the sentence example, the number of capitalized words is 1, the number of hashtags is 1, and the text length is 8.

3. The next step is implementing a sequence-based CNN to extract the local features by sliding the convolutional filters over the token's vector. Suppose the window size is two, and the kernels with size 2*3 are used. Table III illustrates the kernel operations to extract the feature map.

TABLE II.BERT TOKENIZATION

Word	BERT tokenization vector
#	[0.1, 0.1, 0.1]
I	[0.5, 0.6, 0.7]
am	[0.4, 0.5, 0.6]
excited	[0.8, 0.9, 1.0]
to	[0.2, 0.3, 0.4]
join	[0.5, 0.6, 0.7]
you	[0.7, 0.8, 0.9]
tomorrow	[0.4, 0.5, 0.6]

TABLE III. KERNELS OPERATIONS

Token's vector	Kernel1's weights	Kernel1 output
# + I	[[0.2,0.3,0.4], [0.5,0.6.0.7]]	1.19
I + am		1.48
am + excited		2.11
excited + to		1.39
to + join		1.39
join + you		2.02
you + tomorrow		1.66
Token's vector	**Kernel2's weights**	**Kernel2 output**
# + I	[[0.2,0.1,0.3], [0.6,0.4.0.5]]	0.86
I + am		1.11
am + excited		1.65
excited + to		0.99
to + join		1.88
join + you		1.56
you + tomorrow		3.66

After extracting the feature map, which is the kernel1 and kernel2 outputs, the MaxPooling operation highlights the crucial features. In this example, the MaxPooling operation takes the maximum feature value between two. Table IV shows the MaxPooling operation. The selected MaxPooling features for each kernel are transferred to the next step, the attention layer operation, to weigh the importance of each extracted feature. The attention layer operation focuses on emotionally relevant parts that refer to happiness. Table V presents attention layer results, where the results of kernel1 and kernel2 are [1.81] and [2.44], respectively.

4. The output from the sequence-based CNN and M-extracted features is transferred to the VI-CNN layer to merge using the Harmonic mean, and the final weighted fused γ_2 features are illustrated in Table VI.

TABLE IV. MAXPOOLING OPERATION

Kernel number	Kernel's output	MaxPooling
1	[1.19, 1.48, 2.11, 1.39, 1.39, 2.02, 1.66]	[1.48, 2.11, 2.02, 1.66]
2	[0.86, 1.11, 1.65, 0.99, 1.88, 1.56, 3.66]	[1.11, 1.65, 1.88, 3.66]

TABLE V. ATTENTION LAYER RESULTS

Kernel number	exponential operation	Exp sum	SoftMax	Weighted attention output γ_1
1	exp ([1.48, 2.11, 2.02, 1.66])	[4.02, 5.73, 5.49, 4.51] = 19.75	[0.20, 0.29, 0.27, 0.22]	1.48*0.20+2.11*0.29+2.02*0.27+ 1.66*0.22= 1.81
2	exp ([1.11, 1.65, 1.88, 3.66])	[3.01, 4.48, 5.11, 9.94] = 22.63	[0.13, 0.19, 0.22, 0.43]	1.11*0.13+1.65*0.19+1.88*0.22+ 3.66*0.43= 2.44

TABLE VI. FUSED FEATURES

Weighted attention output γ_1	$M_{extracted}$	VI-CNN layer output γ_2
1.81	1	2/(1/1.81+1/1)=2/1.55=1.29
1.81	1	2/(1/1.81+1/1)=1.29
1.81	8	2/(1/1.81+1/8)=2/1.675=1.19
2.44	1	2/(1/2.44+1/1)=2/1.40=1.42
2.44	1	2/(1/2.44+1/1)=1.42
2.44	8	2/(1/2.44+1/8)=2/0.53=3.77

TABLE VII THE VIT OPERATION

Patch number	patches	Attention weight	Exponential operation	SoftMax	Attention output
1	[1.29, 1.29,1.19]	Q1.K1=1.29*1.29+1.29 *1.29+1.19*1.19=5.70	Exp(5.70)=15.49	15.49/37.61=0.41	0.41*[1.29, 1.29, 1.19] + 0.58*[1.42, 1.42, 3.77] = [0.52, 0.52, 0.48] + [0.82, 0.82, 2.18] =[1.34, 1.34, 2.66]
2	[1.42,1.42, 3.77]	Q1.K2=1.29*1.42+1.29 *1.42+1.19*3.77=8.14	Exp(8.14)=22.12	22.12/37.61=0.58	

5. The fused feature vector y2 is input to the ViT model, which divides it into patches. Each patch is flattened and projected into the transformer's embedding space. Then the transformer encoder applies self-attention and feed-forward to learn the contextual relationships using (39)

$$\text{Attention}(Q,K,V) = S\text{oftMax}\left(\frac{QK^T}{\sqrt{dk}}\right)\text{V} \quad 39$$

Q refers to Query, K to Key, and V to Values, and here, they are identical. After calculating the attention output, the dot product operation is applied between it and the fully connected layer weight. Then, the SoftMax operation is used to get the probability for each emotion class, γ_3. Table VII shows the attention mechanism. Suppose the final probability for each class, love is 0.2, sadness is 0.19, happiness is 0.35, and neutral is 0.26. So, this example's final prediction γ_3 is happiness because it has the highest probability.

I. RESULTS AND DISCUSSION

The validation of experimental results attained by Vi-CNN and other baseline approaches employed for comparison, and the discussions performed are as follows.

A. Experimental set-up

The Vi-CNN developed for emotion detection is executed using a Python tool with Keras and TensorFlow libraries.

B. Performance analysis

The effectiveness of Vi-CNN designed for emotion detection is validated for several hidden neurons and the number of epochs by varying learning sets using different evaluation measures.

1) Performance measures

To evaluate the superiority of Vi-CNN used for emotion detection, the following metrics have been used:

- *F-measure:* It is the harmonic mean between precision and recall and is expressed as in (40),

$$F-measure = 2\frac{R*P}{R+P} \quad (40)$$

Here, the recall is signified as *R* and *P* indicates precision.

- *Recall:* Recall is the ratio of accurately detected positive labels to the total positive labels and is given by (41),

$$Rec\,all, R = \frac{u_1}{u_1+u_4} \quad (41)$$

where, true positive is signified as u_1, and u_4 represents false negative.

- *Perecision:* Precision is the ratio of accurately detected positive labels from the total detected positive labels, and is articulated as in (42),

$$Pr\,ecision, P = \frac{u_1}{u_1+u_3} \quad (42)$$

here, false positive is indicated by u_3 [36].

2) Validation of Vi-CNN concerning the number of epochs

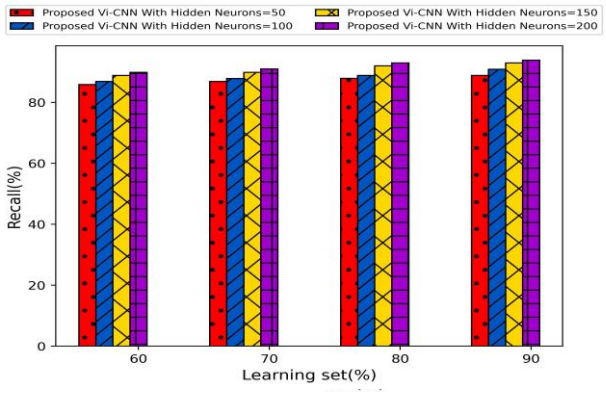
The analysis of the performance of Vi-CNN utilized for detecting emotion from texts based on many epochs is delineated in Fig. 5. The evaluation of Vi-CNN using recall by altering total epochs is demonstrated in Fig. 5(a). For 90% of the learning set, the Vi-CNN measured recall of 93.877%, 91.887%, 89.978%, and 88.876% for 40, 30, 20, and 10 epochs. Furthermore, the validation of Vi-CNN developed for emotion detection from input text utilizing F-measure is elucidated in Fig. 5(b).

Here, for 40, 30, 20, and 10 epochs, the Vi-CNN measured F-measure of 92.867%, 90.882%, 89.480%, and 88.384% as for the 90% learning set. Also, Fig. 5(c) portrays the valuation of Vi-CNN using precision, whereas the precision of 91.878%, 89.898%, 88.988%, and 87.898% is recorded by Vi-CNN for 40, 30, 20, and 10 epochs as for learning a set of 90%.

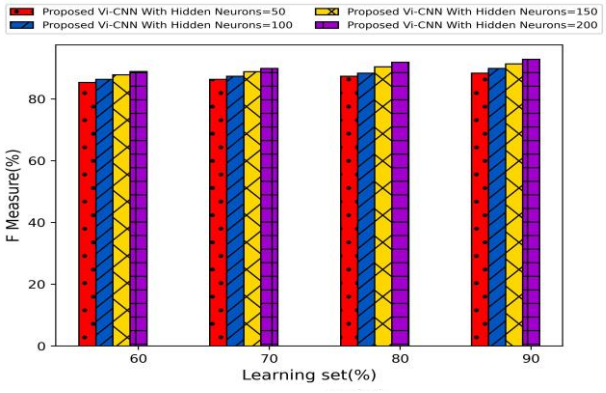
3) Validation of Vi-CNN concerning the number of hidden neurons

The performance valuation of Vi-CNN used for emotion detection from input text based on the total number of neurons is depicted in Fig. 6. The evaluation of Vi-CNN utilizing recall for different numbers of neurons is depicted in Fig. 6(a). The Vi-CNN observed recall of 93.877%, 92.987%, 90.898%, and 88.876% for hidden neurons of 10, 8, 6, and 4 for 90% learning set. Fig. 6(b) demonstrates the valuation of Vi-CNN based on F-measure, where for a learning set of 90%, F-measure of 92.867%, 91.410%, 89.932%, and 88.373% is measured by Vi-CNN with a total of 10, 8, 6, and 4 hidden neurons. Moreover, the evaluation of Vi-CNN utilizing precision for different numbers of neurons is illustrated in Fig

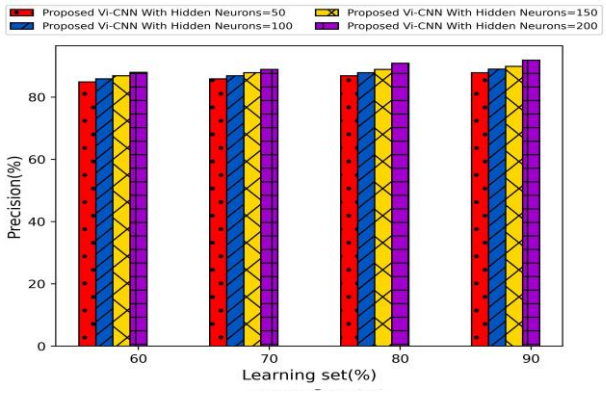
6(c). For 10, 8, 6, and 4 hidden neurons, the Vi-CNN recorded a precision of 91.878%, 89.887%, 88.987%, and 87.876% for the learning set of 90%.



a



b



c

Fig. 6. Performance evaluation of Vi-CNN using (a) Recall, (b) F-measure, and (c) Precision based on the number of neurons

C. Comparative schemes and validation

The prevailing schemes used for emotion detection from text data, such as CSAGAT [20], TA-MERT [11], BiLSTM [21], and DIMMN [8], are compared to identify Vi-CNN's superiority in detecting human emotion. The comparative validation is performed to determine the effectiveness of Vi-CNN by varying the K-set and learning set, where the analysis executed is explicated as follows,

1) Validation of Vi-CNN concerning K-set

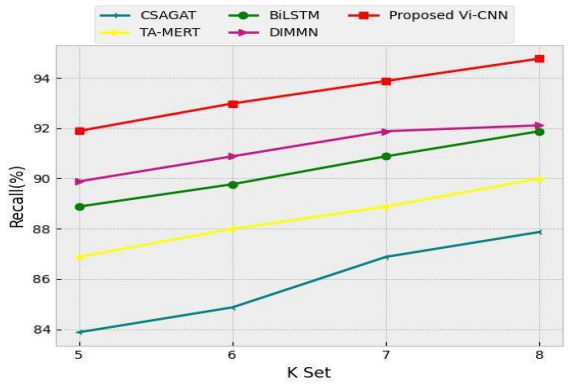
The experimental valuation of Vi-CNN designed for emotion detection in this research by considering the K-set is given in Fig. 7. The validation of Vi-CNN employing recall is demonstrated in Fig. 7(a), where Vi-CNN measured a maximum recall of 94.765% for a K-set of 8. Likewise, the baseline emotion detection schemes recorded recall of 87.868% by CSAGAT, 89.989% by TA-MERT, 91.877% by BiLSTM, and 92.109% by DIMMN. As seen in the figure, the Vi-CNN attained superior results with a maximum performance of 2.80% compared to the baseline DIMMN scheme. In Fig. 7(b), the F-measure observed by Vi-CNN and other prevailing emotion detection approaches employed for comparison is validated by varying the K-set. For the K-set of 8, the baseline detection schemes, like CSAGAT, TA-MERT, BiLSTM, and DIMMN, recorded F-measure of 85.803%, 87.313%, 89.430%, and 90.877%, whereas the Vi-CNN outperforms prevailing models with a maximum F-measure of 91.877%. The results proved that Vi-CNN attained superior results with a higher performance of 4.97% than TA-MERT. Further, the valuation of Vi-CNN designed for detecting emotions from textual data using precision is delineated in Fig. 7(c). The Vi-CNN recorded a maximum precision of 92.988%, and the precision measured by baseline schemes, like CSAGAT, is 86.766%, TA-MERT is 88.879%, BiLSTM is 89.899%, and DIMMN is 90.868% for the K-set of 8. The results revealed that Vi-CNN achieves the maximum performance of 6.69% compared to the traditional CSAGAT model.

2) Validation of Vi-CNN with respect to the learning set

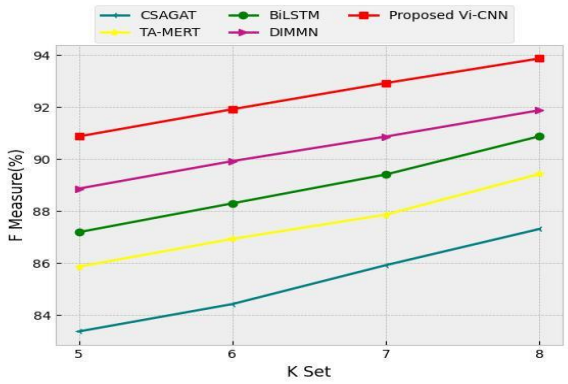
The assessment of Vi-CNN for detecting emotions based on the learning set is elucidated in Fig. 8. The analysis of Vi-CNN developed to detect emotions from textual data, considering recall, is portrayed in Fig. 8(a). The Vi-CNN measured a high recall of 93.878% for 90% of the learning set, and the recall measured by baseline schemes, like CSAGAT, is 85.898%, TA-MERT is 87.988%, BiLSTM is 89.887%, and DIMMN is 90.877%. The Vi-CNN achieved a maximum performance of 3.20% compared to traditional DIMMN. The evaluation of Vi-CNN utilized for emotion detection by utilizing F-measure is depicted in Fig. 8(b). Here, the F-measure of 92.867% is obtained by Vi-CNN for 90% of the learning set. Similarly, the baseline schemes measured an F-measure of 85.440% by CSAGAT, 86.920% by TA-MERT, 88.928% by BiLSTM, and 90.374% by DIMMN. It can be observed that the Vi-CNN attained a superior performance of 4.24% compared to the prevailing BiLSTM model. Fig. 8(c) shows the precision measured by Vi-CNN and existing techniques utilized for emotion detection with varying learning sets.

The existing techniques, such as CSAGAT, TA-MERT, BiLSTM, and DIMMN, measured precision of 84.988%, 85.879%, 87.990%, and 89.876% for a learning set of 90%. The Vi-CNN outperforms baseline schemes used for emotion

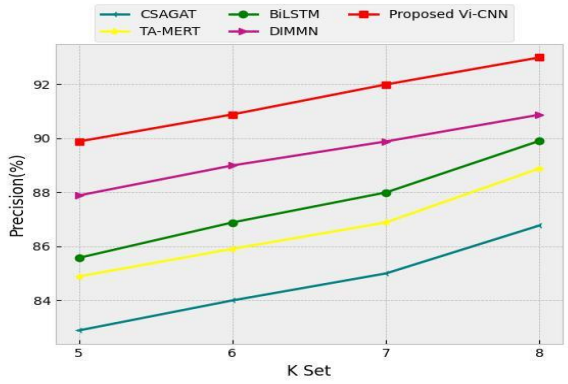
detection with a maximum precision of 91.879%. The results show that Vi-CNN achieved superior experimental outcomes with a higher performance by 7.50% than CSAGAT.



a



b

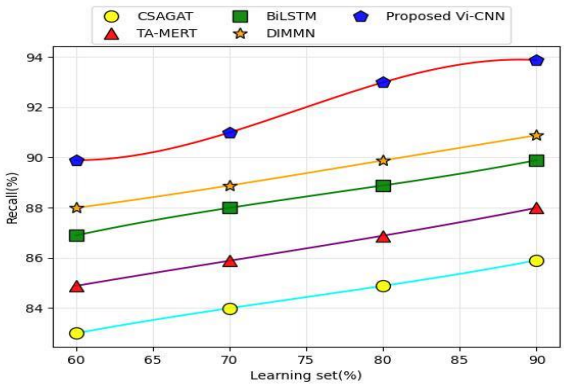


c

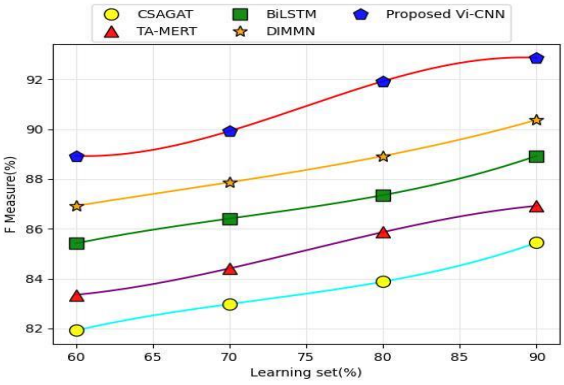
Fig. 7. Comparative validation of Vi-CNN utilizing (a) Recall, (b) F-measure, and (c) Precision concerning K-set

D. Comparative discussion

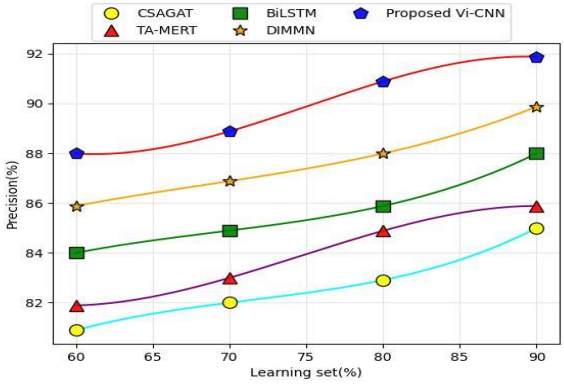
Table VIII illustrates the results obtained from the experiment by Vi-CNN and existing emotion detection schemes concerning the K-set and learning set. The Vi-CNN outperforms existing techniques with a high recall of 94.765%, F-measure of 91.877%, and precision of 92.988% for a K-set of 8. Similarly, the other prevailing approaches attained recall of 87.868% by CSAGAT, 89.989% by TA-MERT, 91.877% by BiLSTM, and 92.109% by DIMMN. The F-measure recorded by existing schemes, like CSAGAT, TA-MERT, BiLSTM, and DIMMN, is 86.766%, 88.879%, 89.899%, and 90.868%, and the existing techniques measured precision of 85.803%, 87.313%, 89.430%, and 90.877%. The results show that the Vi-CNN precisely detected emotions from the text samples to test the model's performance. The Vi-CNN classified emotional states from many sequential text data to recognize emotional patterns. Also, the Vi-CNN significantly reduced the complexity of the models, thereby increasing the generalization ability to enhance the detection performance.



a



b



c

Fig. 8. Comparative validation of Vi-CNN employing (a) Recall, (b) F-measure, and (c) Precision concerning the learning set

TABLE VIII. COMPARATIVE DISCUSSION

Variations	Metrics (%)	CSAGAT	TA-MERT	BiLSTM	DIMMN	Designed Vi-CNN
For learning set 90%	***Recall***	85.898	87.988	89.887	90.877	93.878
	Precision	84.988	85.879	87.990	89.876	91.879
	F-measure	85.440	86.920	88.928	90.374	92.867
For K-set 8	***Recall***	87.868	89.989	91.877	92.109	***94.765***
	Precision	86.766	88.879	89.899	90.868	***92.988***
	F-measure	87.313	89.430	90.877	91.876	***93.867***

In the field of emotion recognition, sometimes text may express multiple emotions. For instance, "I'm happy but also a bit nervous about today" could hold both happy and worry feelings. The VI-CNN system integrates multiple mechanisms to obtain nuanced and combined emotional features. Accordingly, the extracted local features by CNN are fused with engineered features, yielding a rich feature set. In addition to the attention mechanism, the system can notice multiple emotionally salient words. Finally, using SoftMax probabilities enables the system to express that more than one emotion is relevant to an input.

The proposed Vi-CNN model outperforms other models as the convolution operation effectively captures local patterns and context, regardless of their position in the input. This helps robustly identify critical emotion-related features (e.g., specific words or visual cues), even if they appear at different positions.

Additionally, the model utilizes the vision sequence-based CNNs that maintain sequential information relying on the spatial arrangement of embedding. The vision sequence can effectively capture local sequential dependencies through convolution and pooling operations. Therefore, when the proposed paradigm is contrasted with pure transformer models having an exclusive reliance on global self-attention mechanisms, it is apparent that convolutional neural networks or CNNs can provide a more computationally efficient approach while still having the ability to maintain the potential for capturing the intrinsic structure of sequences needed for understanding the data. Lastly, Vision sequence-based CNNs can naturally combine text embedding with other text features, such as punctuation and capitalization, in a spatially coherent way, in contrast to different models that treat each set of features separately.

II. CONCLUSION

Nowadays, emotion detection has gained massive attention from humans due to increasing advancements in identifying individual opinions, emotions, and sentiments from textual conversations. Various techniques are used to accurately detect the individual's emotions from textual data to understand the user experiences, communication, and personalization. In this article, a hybrid deep learning model, namely Vi-CNN, is designed for emotion detection from text data. Here, the emotions are accurately detected using Vi-CNN and classified into anger, boredom, enthusiasm, emptiness, fun, sadness, neutral, hate, relief, love, happiness, surprise, and worry. The supremacy of Vi-CNN in emotion detection is validated, where the Vi-CNN attained superior results with F-measure, precision, and recall of 93.867%, 92.988%, and 94.765%. In future research, hybrid algorithmic techniques will be employed to optimally adjust the hyperparameters of the deep learning approach to enhance performance.

REFERENCES

[1] Saxena, A., Khanna, A. and Gupta, D., "Emotion recognition and detection methods: A comprehensive survey", Journal of Artificial Intelligence and Systems, vol.2, no.1, pp.53-79, 2020 doi: https://doi.org/10.33969/AIS.2020.21005.

[2] Khare, S.K., Blanes-Vidal, V., Nadimi, E.S. and Acharya, U.R., "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations", Information Fusion, vol.102, pp.102019, 2024. doi:https://doi.org/10.1016/j.inffus.2023.102019.

[3] Torres-Valencia, C., Álvarez-López, M. and Orozco-Gutiérrez, A., "SVM-based feature selection methods for emotion recognition from multimodal data", Journal on Multimodal User Interfaces, vol.11, pp.9-23, 2017. https://doi.org/10.1007/s12193-016-0222-y

[4] Abas, A.R., Elhenawy, I., Zidan, M. and Othman, M., "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text", Computers, Materials & Continua,

vol.71, no.2, 2022. https://doi.org/10.32604/cmc.2022.021671

[5] Halim, Z., Waqar, M. and Tahir, M., "A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email", Knowledge-based systems, vol.208, pp.106443, 2020. https://doi.org/10.1016/j.knosys.2020.106443

[6] Doewes, R.I., Gangadhar, L. and Subburaj, S., "An overview on stress neurobiology: Fundamental concepts and its consequences", Neuroscience Informatics, vol.1, no.3, pp.100011, 2021, doi:https://doi.org/10.1016/j.neuri.2021.100011

[7] Shelke, N., Chaudhury, S., Chakrabarti, S., Bangare, S.L., Yogapriya, G. and Pandey, P., "An efficient way of text-based emotion analysis from social media using LRA-DNN", Neuroscience Informatics, vol.2, no.3, pp.100048, 2022, doi:https://doi.org/10.1016/j.neuri.2022.100048

[8] Wen, J., Jiang, D., Tu, G., Liu, C. and Cambria, E., "Dynamic interactive multiview memory network for emotion recognition in conversation", Information Fusion, vol.91, pp.123-133, 2023, doi:https://doi.org/10.1016/j.inffus.2022.10.009

[9] Adikari, A., Gamage, G., De Silva, D., Mills, N., Wong, S.M.J. and Alahakoon, D., "A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web", Future Generation Computer Systems, vol.116, pp.302-315, 2021, doi:https://doi.org/10.1016/j.future.2020.10.028

[10] Jianqiang, Z., Xiaolin, G. and Xuejun, Z., "Deep convolution neural networks for twitter sentiment analysis", IEEE access, vol.6, pp.23253-23260, 2018. doi: https://doi.org/10.35940/ijitee.i1107.0789s419.

[11] Mohammad, F., Khan, M., Marwat, S.N.K., Jan, N., Gohar, N., Bilal, M. and Al-Rasheed, A., "Text augmentation-based model for emotion recognition using transformers", CMC-COMPUTERS MATERIALS & CONTINUA, vol.76, no.3, pp.3523-3547, 2023, doi: https://doi.org/10.32604/cmc.2023.040202.

[12] Batbaatar, E., Li, M. and Ryu, K.H., "Semantic-emotion neural network for emotion recognition from text", IEEE access, vol.7, pp.111866-111878, 2019. doi: https://doi.org/10.1109/ACCESS.2019.2934529.

[13] Han, T., Zhang, Z., Ren, M., Dong, C., Jiang, X. and Zhuang, Q., "Text emotion recognition based on XLNet-BiGRU-Att", Electronics, vol.12, no.12, pp.2704, 2023, doi: https://doi.org/10.3390/electronics12122704

[14] Sailunaz, K. and Alhajj, R., "Emotion and sentiment analysis from Twitter text", Journal of computational science, vol.36, pp.101003, 2019, doi: https://doi.org/10.1016/j.jocs.2019.05.009

[15] Mohammad, S.M. and Bravo-Marquez, F., "WASSA-2017 shared task on emotion intensity", arXiv preprint arXiv:1708.03700, 2017, doi:https://doi.org/10.48550/arXiv.1708.03700

[16] Bharti, S.K., Varadhaganapathy, S., Gupta, R.K., Shukla, P.K., Bouye, M., Hingaa, S.K. and Mahmoud, A., "Text-Based Emotion Recognition Using Deep Learning Approach", Computational Intelligence and Neuroscience, vol.2022, no.1, pp.2645381, 2022, doi: https://doi.org/10.1155/2022/2645381

[17] Rastogi, G. and Sushil, R., "Cloud computing implementation: key issues and solution", In proceedings of 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom, pp. 320-324, March, 2015.

[18] Irfan, S., Anjum, N., Althobaiti, T., Alotaibi, A.A., Siddiqui, A.B. and Ramzan, N., "Heartbeat classification and arrhythmia detection using a multi-model deep-learning technique", Sensors, vol.22, no.15, pp.5606, 2022, doi:https://doi.org/10.3390/s22155606

[19] Wani, T.M., Gunawan, T.S., Qadri, S.A.A., Kartiwi, M. and Ambikairajah, E., "A comprehensive review of speech emotion recognition systems", IEEE access, vol.9, pp.47795-47814, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3068045

[20] Tu, G., Wen, J., Liu, C., Jiang, D. and Cambria, E., "Context-and sentiment-aware networks for emotion recognition in conversation", IEEE Transactions on Artificial Intelligence, vol.3, no.5, pp.699-708, 2022, doi: https://doi.org/10.1109/TAI.2022.3149234

[21] Asghar, M.Z., Lajis, A., Alam, M.M., Rahmat, M.K., Nasir, H.M., Ahmad, H., Al-Rakhami, M.S., Al-Amri, A. and Albogamy, F.R., "A deep neural network model for the detection and classification of emotions from textual content", Complexity, vol.2022, no.1, pp.8221121, 2022, doi: https://doi.org/10.1155/2022/8221121

[22] Ghafoor, Y., Jinping, S., Calderon, F.H., Huang, Y.H., Chen, K.T. and Chen, Y.S., "TERMS: textual emotion recognition in multidimensional space", Applied Intelligence, vol.53, no.3, pp.2673-2693, 2023, doi:https://doi.org/10.1007/s10489-022-03567-4

[23] Emotion detection from text dataset is taken from "https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text?resource=download" accessed on November 2024.

[24] Grail, Q., Perez, J. and Gaussier, E., "Globalizing BERT-based transformer architectures for long document summarization", In Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume, pp. 1792-1810, 2021, doi:https://doi.org/10.18653/v1/2021.eacl-main.154

[25] Sundaram, V., Ahmed, S., Muqtadeer, S.A. and Reddy, R.R., "Emotion analysis in text using TF-IDF", In proceedings of 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 292-297, January, 2021, doi:https://doi.org/10.1109/Confluence51648.2021.9377159

[26] Thakur, R.K. and Deshpande, M.V., "Kernel Optimized-Support Vector Machine and Mapreduce framework for sentiment classification of train reviews", International

Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.27, no.06, pp.1025-1050, 2019, doi:https://doi.org/10.1142/S0218488519500454

[27] Gupta, A. and Goyal, K.K., "Classification of Semantic Similarity Technique between Word Pairs using Word Net", International Journal of Engineering and Advanced Technology, vol.9, No.2, pp.4397-4402, 2019.

[28] Shrivastava, K., Kumar, S. and Jain, D.K., "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network", Multimedia tools and applications, vol.78, pp.29607-29639, 2019, doi:https://doi.org/10.1007/s11042-019-07813-9

[29] Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J. and Molchanov, P., "A-vit: Adaptive tokens for efficient vision transformer", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10809-10818, 2022, doi: https://doi.org/10.1109/cvpr52688.2022.01054.

[30] Suanmali, L., Salim, N. and Binwahlan, M.S., "Feature-based sentence extraction using fuzzy inference rules", In Proceedings of 2009 International Conference on Signal Processing Systems, IEEE, pp. 511-515, 2009, doi:https://doi.org/10.1109/ICSPS.2009.156

[31] Ghosh, M. and Kar, A., "Unsupervised linguistic approach for sentiment classification from online reviews using SentiWordNet 3.0", Int J Eng Res Technol, vol.2, no.9, pp.1-6, 2013,

[32] Abubakar, H.D., Umar, M. and Bakale, M.A., "Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec&quot", SLU Journal of Science and Technology, vol.4, no.1, pp.27-33, 2022, doi:https://doi.org/10.56471/slujst.v4i.266

[33] Damsleth, E. and Spjøtvoll, E., "Estimation of trigonometric components in time series", Journal of the American Statistical Association, vol.77, no.378, pp.381-387, 1982, doi:https://doi.org/10.1080/01621459.1982.10477820

[34] J. G. Proakis and D. G. Manolakis, Digital signal processing : principles, algorithms, and applications. Upper Saddle River, N.J.: Prentice Hall, 1996.

[35] A. V. Oppenheim and R. W. Schafer, Discrete-time signal processing. Upper Saddle River: Pearson, 2010.

[36] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," Transactions of the Association for Computational Linguistics, vol. 12, pp. 820–836, Jan. 2024, doi: https://doi.org/10.1162/tacl_a_00675.