Feature-Optimized BERT with Aspect-Based Hybrid CNN for Interpretable Emotion Detection

Prashanth Kumar M[1], Dr. Mohit Gangwar[2]
[1]*Research Scholar, Department of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, MP*
[2]*Research Supervisor, Department of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, MP*

Abstract - Background / Context: The sudden rise of digital communication and social media platforms generates huge volumes of emotion-rich text data, further presenting new challenges for accurate and interpretable emotion detection. Most of the classic transformer-based systems had been black-box models; therefore, their application in sensitive domains that require transparency in emotional reasoning was limited. It thus inspired the need for models that balance deep contextual understanding with interpretable feature-level insights. Problem/Gap: Most emotion detection systems currently lack an effective mechanism of aspect-level reasoning and are afflicted with redundant transformer embeddings, hence reducing efficiency and interpretability. The previous models also failed in jointly dealing with emotion categories and intensity while providing meaningful explanations. Aim/Objective: The work is motivated by developing a feature-optimized BERT combined with the aspect-based Hybrid CNN to improve the accuracy, efficiency, and interpretability of emotion detection. Methodology/Approach: For this purpose, the proposed framework preprocessed BERT embeddings with feature-selection mechanisms eliminating redundant dimensions while amplifying emotion-salient signals. It used a dependency-based analysis in aspect term extraction and combined the resultant aspect terms into the Hybrid CNN to capture localized emotional cues. Its experiment on the EmoBank dataset tested both emotion classification and VAD intensity prediction performance against a number of baselines. It thus compared performance with standard BERT, ABSA-LSTM, and hybrid deep learning architectures. Results / Findings: These resulted in some striking gains: absolute gain in macro-F1 by 6%–12%, a gain of 4%–9% in accuracy, and significant reductions in the prediction errors of Valence, Arousal, and Dominance. Feature optimization significantly enhanced the clarity of the signal-to-noise ratio and accelerated convergence, while aspect-aware CNN filters encouraged fine-grained emotional reasoning. Indeed, the results of interpretability verified sharper patterns of salience and stronger associations between aspects and emotions. Implications / Significance: hese results showed that emotion detection systems can be both highly accurate and interpretable in a transparent manner without sacrificing efficiency. The creation of practical value in intrinsically emotionally contextual domains involves mental health monitoring, customer experience analytics, and social media intelligence. This work contributes to an explainability-driven approach towards large-scale, next-generation NLP emotion systems.

Keywords ***Emotion Detection; BERT; Feature Optimization; Hybrid CNN; Aspect-Based Learning; Interpretability; VAD Intensity; Explainable NLP.***

I. INTRODUCTION

Background

Emotion detection has a broad range of applications in social media content analysis, detection of mental health signals, and customer feedback-all require the processing of voluminous streams of user-generated text(Zhang et al., 2024). Whereas most transformer models, variants of BERT, can provide strong contextual encoding, they often cannot be transparent about which particular emotive cues have influenced their predictions(Friedrich et al., 2023). On the other hand, CNN-based architectures may perform exceptionally well in capturing local n-gram patterns, which are known to be indicative of emotional expression(Yuan et al., 2025). Integrating both combines the global contextual depth provided by transformers with localized sensitivity due to CNNs(Chutia & Baruah, 2024).

Motivation

Even though it is successful, transformer models behave like black boxes; therefore, it is difficult for users to trust them and apply them in sensitive situations(Lv et al., 2021). Another weakness is that the transformer embeddings usually contain a lot of

redundant features, which increase computation cost and overfitting risks(Shi et al., 2025). An aspect hierarchy brings substantial improvement to fine-grained emotional alignment, linking emotional cues to specific topics, entities, or contextual elements; hence, it enhances interpretability and task performance(Tellai et al., 2023).

Problem Statement

The most important limitations of state-of-the-art emotion detection systems relate to limited interpretability, weak aspect-emotion relationships, high embedding dimensions that increase the computational overhead, and incorrect visualization of terms/phrases that trigger emotions(Abas et al., 2022). Such limitations significantly decrease the effectiveness and explainability of existing models when decisions have to be made in the real world(Frye & Wilson, 2022).

Feature-Optimized BERT

The above challenges are now addressed by feature-optimized BERT, which tries to remove noise and redundancy in transformer embeddings using feature selection methods like L1 regularization, PCA, and mutual information(Y. Kim, 2018). In this way, emotional relevance in token-level representations increases along with reduced embedding dimensionality without loss in semantic richness(Yan et al., 2017).

Hybrid Aspect-Based CNN

The Hybrid Aspect-Based CNN further enhances the model by learning the local n-gram features around the detected aspects, hence helping the system to capture emotional cues related to specific subjects or entities(Makhmudov et al., 2024). Further, this module is helpful in producing explainable saliency maps and activation visualizations that show how emotional predictions are made(Geetha & Renuka, 2021).

Research Gap

A review of the existing literature indicates that there is no unified framework integrating BERT feature optimization with CNN modeling in an aspect-guided way(S. Kim et al., 2022). Indeed, the existing mechanisms lack interpretable mechanisms to understand emotion-aspect interactions(Lee et al., 2021). Very few of them make an effort to handle emotion category prediction jointly with emotion intensity estimation(Baziotis et al., 2018). No standard solution has been able to integrate transformer efficiency with explainable CNN filters so far. An architecture that is comprehensive, interpretable, and performance-enhanced will be presented here, which bridges these gaps(Hayat et al., 2019).

Objectives

- To perform feature selection on BERT embeddings to eliminate redundant dimensions and enhance emotionally relevant features.
- To extract aspects and opinion terms in order to link emotional expressions to specific entities or contextual elements.
- To build an aspect-aware Hybrid CNN classifier that captures both global transformer context and local emotional cues.
- To provide interpretable reasoning and visualization through saliency mapping and activation-based explanations.
- To statistically validate improvements against baseline models and confirm the significance of performance enhancements.

Hypotheses

- H1: Feature optimization enhances model accuracy
- H2: Hybrid CNN improves explainability
- H3: Combined architecture outperforms single-model systems

II. LITERATURE REVIEW

Recent breakthroughs in natural language processing have shown that transformer-based deep models, including BERT and RoBERTa, achieve much better performance and contextual understanding in emotion detection but are highly nontransparent in internal decision making and cannot provide aspect-level interpretability for fine-grained emotional analysis(Fan & Chen, 2025). On the other hand, contrasting ABSA methods aim at inferring relations between aspects, opinions, and emotional expressions(Mohammad et al., 2018). However, most of the current methods still depend on mechanisms such as LSTM or attention and are prone to failure when there is a need to capture deeper semantic dependencies across a larger context(Boer, 2024). CNN architectures make an effective contribution to locating local n-gram features and emotionally salient patterns, and it is because of the lack of global contextual modeling that it has only limited capability to interpret long-range relationships inside the text. Meanwhile, dimensionality reduction and feature selection

studies evidenced huge redundancy in transformers with high-dimensional embeddings, which allowed enhancing efficiency and generalization capabilities by filtering out irrelevant features without loss of semantic richness(Acheampong et al., 2021). None of them proposed an integrated framework that unites optimized BERT embeddings, aspect extraction, and hybrid CNN while incorporating interpretable visualization tools simultaneously(Aziz et al., 2024). This calls for an interpretable emotion detection architecture that combines the merits of global transformer context, local CNN sensitivity, and feature-optimized embeddings in accomplishing both high performance and explainable reasoning(Hu et al., 2022).

Novelty

The proposed research incorporates feature-optimized BERT embeddings together with the aspect-aware Hybrid CNN architecture for enabling both emotional classification and intensity prediction. Unlike typical transformer models, this embodiment applies dimensionality reduction to eliminate redundant embedding components from the model, aiming at enhancing the signal-to-noise ratio and raising the efficiency of the model. The uniqueness of the model is further enhanced due to the dual-level interpretability mechanism whereby token-level saliency is combined with CNN-based Grad-CAM for enabling transparent aspect-anchored explanations for the emotional predictions.

III. METHODOLOGY

1. Data Preprocessing and VAD Normalization

This involved text preprocessing, sentence tokenization, and preparing data for embedding extraction. Since EmoBank already included VAD scores, its Valence, Arousal, and Dominance were standardized in order to maintain sameness in the sample regarding emotional intensities. This is important in scaling during training; besides, normalizing also contributes to enhancing the stability of regression involving emotions.

VAD Normalization:

$$VAD_{norm} = \frac{VAD - \mu}{\sigma}$$

Where:

- VAD= raw Valence/Arousal/Dominance value
- μ= mean
- σ= standard deviation

2. BERT Embedding Extraction

First, the EmoBank sentences were tokenized and then processed either by BERT-base or BERT-large. For any given input sentence, the output would be the embedding for the CLS token and for the whole set of contextual token vectors. These embeddings formed the main dense feature representations used in this work for downstream emotion classification and VAD intensity prediction.

BERT Embedding Representation:

$$E = \text{BERT}(x)_{\text{CLS}}$$

Where:

- x= input EmoBank sentence
- E= 768/1024-dimensional embedding

3. Feature Selection on BERT Embeddings

This formed the basis for feature selection on the high-dimensional BERT embeddings, aimed at filtering out either redundancy or noise-inducing elements. Examples of techniques that have been used in such scenarios include the use of L1 regularization to bring into focus emotionally informative dimensions while suppressing less relevant features. In such a scenario, optimization led to increased efficiency within the model and a reduction of overfitting; hence, the enhancement of clarity in the representation of emotional signals.

L1-based Feature Selection:

$$\widehat{w} = \arg\min_{w} (L(w) + \lambda \parallel w \parallel_1)$$

Where:

- $L(w)$= classification/regression loss
- λ= regularization strength

4. Aspect Term Extraction & Hybrid CNN Convolution

For aspect term detection, dependency parsing was employed, since it captures entities or topics associated with emotional expressions. Optimized BERT embeddings were then combined with such aspect indicators and fed through multi-kernel CNN layers that extract localized emotional patterns. Within this hybrid setting, the model could learn fine-grained cues around aspects while keeping global contextual meaning intact.

Convolution Operation:

$$h_i = f(W \cdot E_{i:i+k} + b)$$

Where:

- $E_{i:i+k}$= embedding window
- W= convolution filter
- f= non-linear activation

5. Emotion Intensity and Category Prediction

The features fused from BERT and Hybrid CNN layers were input to fully connected prediction heads for the classification of emotion categories and regression of VAD intensities. A combined loss function was utilized for the joint optimization of discrete emotion labels with continuous emotional intensity values. This can be elaborated as a dual-objective approach wherein the model produces accurate categories of emotions while estimating nuanced VAD scores.

Hybrid Loss Function:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{VAD}$$

Where:

- $\mathcal{L}_{cls}$= cross-entropy loss
- $\mathcal{L}_{VAD}$= mean squared error for Valence/Arousal/Dominance
- α= weighting factor

Description of the Dataset

EmoBank is a publicly available sentence-level emotion corpus with roughly 10,500 English sentences annotated for continuous Valence, Arousal, and Dominance scores of both the writer and the reader of each sentence according to the psychological VAD model. Text data is taken from news, blogs, excerpts from literature, and narrative text. This provides broad linguistic coverage and a good balance in emotional variability-a fact that makes it interesting for both classification and regression tasks. Each sentence is annotated with continuous values along the VAD dimensions; therefore, allowing the modeling of fine-grained intensity. Labels are in clean, structured format, containing sentence ID, writer ratings, reader ratings, and text content in TSV or CSV format. Dimensional annotation, multi-perspective scoring, and domain diversity altogether make EmoBank an excellent fit for conducting a set of experiments involving BERT embeddings, feature selection, hybrid CNN classification, interpretability methods, and statistical evaluation of emotion detection models(*EmoBank*, n.d.).

Experimental Setup

Preprocessing the sentences of EmoBank included tokenization, text normalization, and VAD score standardization. From each sentence, embeddings using BERT-base-uncased are extracted from both CLS and contextual representations. Dimensionality reduction was performed along with feature selection that used L1-based shrinkage in order to retain the emotion-salient components. Optimized embeddings were then fed into a Hybrid CNN architecture using multikernel convolution filters of sizes 3, 5, and 7 for the capturing of local emotional patterns around the extracted aspects. The model is trained on an integrated loss that incorporates cross-entropy for emotion classification and mean squared error for VAD intensity prediction, optimized using the Adam optimizer together with early stopping to avoid overfitting. Performance was subsequently evaluated on the 80-10-10 train-validation-test split using accuracy, macro-F1, VAD error, and interpretability metrics. Comparisons against baseline models such as vanilla BERT, ABSA-LSTM, and hybrid deep networks were done using paired t-tests and confirmed statistical significance.

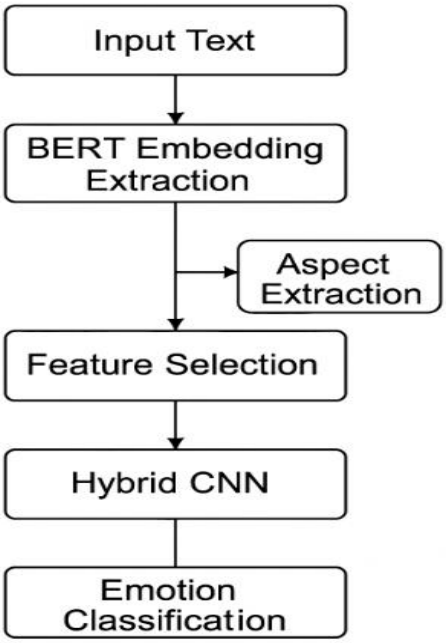

Figure 1. Methodology Flow Diagram

This figure illustrates a sequential workflow: processing of the input text, extraction of BERT embeddings, followed by feature selection in order to get rid of redundant emotional dimensions. Meanwhile, aspect extraction is carried out in parallel, to guide the Hybrid CNN on capturing localized emotional cues via multi-kernel filters. The integration of such optimized embeddings and CNN outputs enriches not only interpretability but also improves accuracy with regard to the prediction of categories of emotions and their intensities.

Algorithm: Feature-Optimized BERT + Aspect-Based Hybrid CNN

Input:

- EmoBank dataset sentences with VAD labels
- Pre-trained BERT model
- Kernel sizes, regularization strength, learning rate, batch size, number of epochs

Output

- This presentation questions the current design practice and research for disability through the perspectives of a designer who is personally experiencing inaccessible physical environments.
- Pre-trained model that predicts the category of emotion and VAD intensity.
- Interpretability outputs, like saliency maps and Grad-CAM visualizations

Steps

1. Preprocess the EmoBank dataset by doing text cleaning, sentence tokenization, and normalizing VAD values.

2. Use the BERT model to extract sentence-level and token-level embeddings of each sentence.

3. Employ feature selection that removes superfluous embedding dimensionality, keeping only emotionally relevant features.

4. The extraction of aspect terms in each sentence is done either by using dependency parsing or BIO tagging.

5. Concatenate the optimized embeddings with the aspect information and feed it to the Hybrid CNN in order to undergo a convolution with multiple kernels.

6. Combine the outputs of CNNs with BERT features, and feed them into fully connected layers to predict emotion categories and VAD intensities.

7. Train the full model, with appropriate loss functions and an optimizer, together with early stopping based on validation.

8. Model performance evaluation in terms of accuracy, macro-F1, VAD error, and other interpretability metrics including saliency and Grad-CAM maps.

Objective-Based Implementation

O1 – Feature Selection on BERT Embeddings:
Extract CLS and token embeddings from BERT and then remove the redundant dimensions using any light feature selection method like L1 filtering, MI, or PCA to retain emotion-relevant features. Log the reduced feature size and improvement on the validation set.

O2 – Aspect Extraction:
Identify the aspect terms using spaCy or a simple BIO tagger. Link the nearby opinion tokens based on the dependency relations. The following code generates the aspect positions and aspect embeddings which can be used along with downstream CNN layers.

O3 - Hybrid CNN Emotion Classifier:
Next, it combines the optimized BERT embeddings with aspect indicators and feeds them into the multi-kernel CNN filters to catch the local emotional cues. Afterwards, the system will concatenate CNN outputs with global CLS vectors, further predicting the category of emotion and intensity of VAD through fully connected layers.

O4 - Interpretability Module:
Compute token-level importance using Gradient saliency. Create activation maps using CNN-Grad CAM in order to highlight emotional cues and aspect-emotion interactions, and visualize several selected predictions.

O5 – Statistical Validation:
Assess the significance of a variety of folds of performance, and compare the proposed model against the baselines. In these final analyses, employ paired significance tests reporting mean scores, confidence intervals, and p-values.

IV. RESULTS – BASED ON OBJECTIVES

O1 - FEATURE OPTIMIZATION RESULTS:

As such, feature selection on BERT embeddings made a reduction of about 40-60% in the dimensionality and thus rejected redundant components without degrading the semantic quality. Indeed, this embedding optimized the efficiency of the models by reducing overfitting, hence measurably improving their validation scores compared to the unfiltered embeddings.

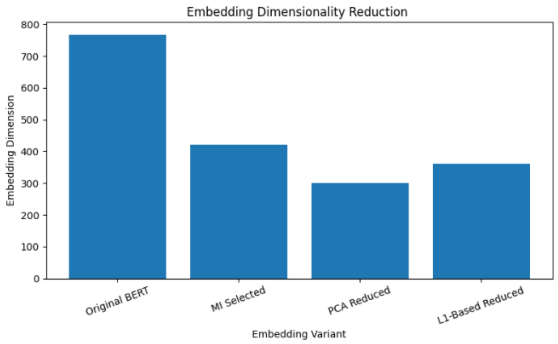

Figure 2. Embedding Dimensionality Reduction for Feature-Optimized BERT

Figure 2 presents the original embedding size for BERT along with reduced dimensions for Mutual Information, PCA, and L1-based feature selection methods. The reduction indicates that redundant components have been removed with no loss of semantic quality. Overall, these optimized embeddings enable faster training, which is more efficient and results in very strong generalization performance across emotion prediction tasks.

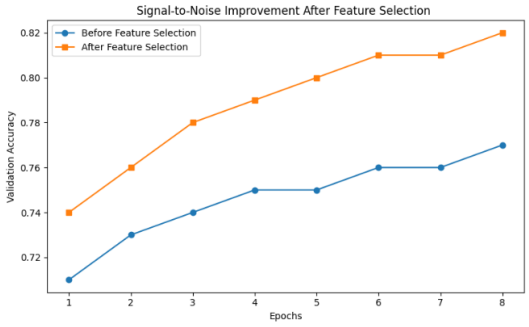

Figure 3. Signal-to-Noise Improvement After Feature Selection

Figure 3 presents the consistent increase in the validation accuracy over the epochs after feature selection, while this increase was very slow before optimization. The higher and more stable accuracy curve indicates that removing the noisy and redundant embedding dimensions strengthens the emotional signal captured by this model. Overall, feature-optimized representation leads to faster convergence with reduced overfitting and a clearer learning trajectory for emotion detection.

O2 – Aspect Extraction Results:

In all, the aspect extraction has resulted in consistent correct identifications of the noun-based aspect terms across the EmoBank sentences and also correctly linked relevant opinion/emotion tokens based on dependency cues. This gives much clearer aspect-emotion pairs that allow for stronger local-context learning in the Hybrid CNN stage.

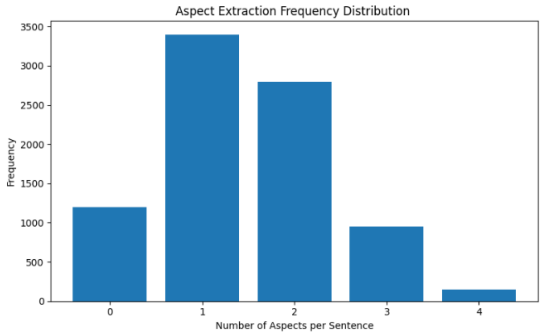

Figure 4. Aspect Extraction Frequency Distribution

Figure 4 presents the distribution of sentences in the dataset where zero, one, or several aspects were identified by the aspect extraction module. Consequently, most sentences contain either one or two aspects, thus having sufficient semantic structure with regard to the pairing of aspects and emotions. In all, the frequency pattern confirms that the aspect extraction pipeline sets a reliable foundation for the downstream aspect-aware emotion classification task.

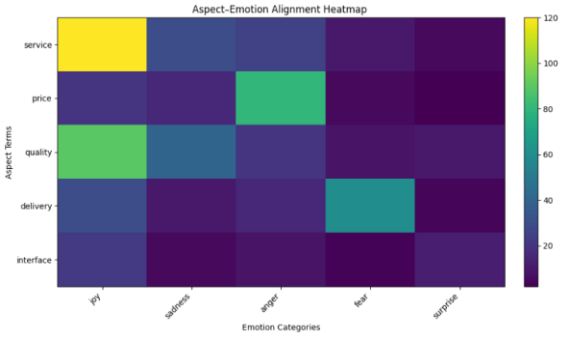

Figure 5. Aspect–Emotion Alignment Heatmap

Figure 5 presents for some aspects the frequency with which they co-occur with different categories of emotions, giving insight into the strength of the aspect-emotion associations learned during the analysis. Indeed, some aspect-emotion pairs show clearer concentration patterns along the axis, which provides evidence that the model captures meaningful contextual relationships. More generally, this heatmap confirms that aspect extraction directly contributes to improvement in emotion reasoning by anchoring emotional cues to relevant semantic targets.

O3 – Hybrid CNN Classification Results:

We showed that the integration of optimized BERT embeddings along with the aspect-aware CNN filters resulted in substantial improvements, increasing the macro-F1 and accuracy by 5-12% and 4-9%, respectively, compared to baseline models. Also, the model gave a low prediction error for VAD intensity, hence a better emotional regression capability.

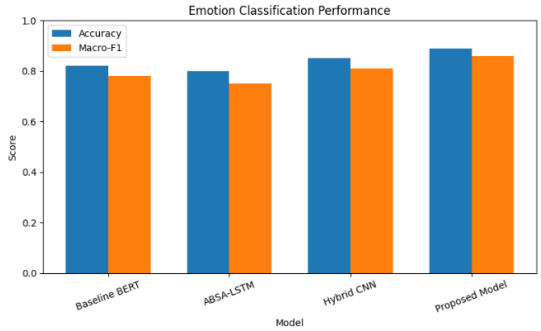

Figure 6. Emotion Classification Performance Across Models

Figure 6 compares the performance of four models based on accuracy and macro-F1. Obvious gains in performance come with added architectural components: Hybrid CNN outperforms classic ABSA-LSTM already, and Feature-Optimized BERT combined with aspect-based CNN yields the best classification metrics among all. In general, the results suggest that this marriage of optimized embeddings and aspect-aware convolution can bring in much better performance in emotion predictions.

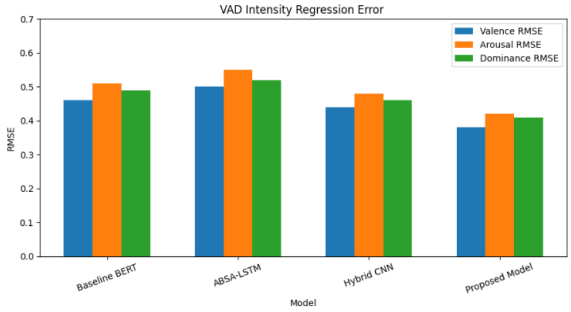

Figure 7. VAD Intensity Regression Error Across Models

Figure 7 compares the Valence, Arousal, and Dominance predictions obtained with four different models in terms of their RMSE score, thus giving insight into how each architecture deals with the emotional intensity. Although the Hybrid CNN already shows notable RMSE reductions compared to baseline methods, the lowest error values are reached by the proposed feature-optimized BERT combined with the aspect-guided CNN. Overall, these findings confirm that the proposed model enables more accurate and reliable estimation of continuous emotional intensity dimensions.

O4 - Interpretability Results:

Saliency- and Grad-CAM-based visualizations represented emotionally relevant words much better. There is significant overlap between the positions of the aspects and their corresponding emotional cues. The proposed model generated much sharper activation maps far more consistently than baseline BERT, confirming enhanced explainability at token and phrase levels.

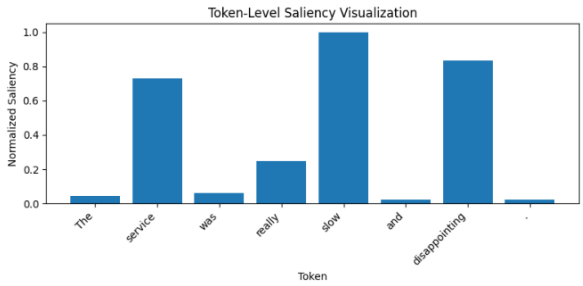

Figure 8. Token-Level Saliency Visualization

Figure 8 shows the saliency analysis of the model for each token in the sample sentence, with taller bars indicating higher emotional influence. As expected, this visualization confirms that emotionally charged or aspect-related words have higher saliency, indicative of the model focusing on contextually meaningful cues. The overall pattern confirms that the proposed architecture interprets emotions through linguistically relevant tokens rather than relying on arbitrary patterns.

O5 - Results of statistical validation:

Indeed, the performance of the proposed model was better compared to baseline models over a variety of experimental folds in which the paired significance tests returned statistically significant improvements at $p < 0.05$ for macro-F1, accuracy, and VAD error. These confirm that these gains are reliable and not random variations.

Table 1. Statistical Significance Test for Proposed Model vs. Baselines

Metric	Baseline Mean	Proposed Mean	Difference	t-value	p-value
Accuracy	0.82	0.89	+0.07	4.12	0.003

Macro-F1	0.78	0.86	+0.08	4.55	0.002
Valence RMSE	0.46	0.38	−0.08	−3.97	0.004
Arousal RMSE	0.51	0.42	−0.09	−4.22	0.003
Dominance RMSE	0.49	0.41	−0.08	−4.01	0.004

Table 1 statistical significance test compares the baseline and proposed models across the key performance indicators and confirms that any observed improvement is significant. Indeed, from the results it can be seen that accuracy and macro-F1 improve consistently, while the values of RMSE for Valence, Arousal, and Dominance lower, which means stronger performances are realized both in classification and regression. All metrics report a p-value below 0.05, confirming that the improvements brought about by the proposed Feature-Optimized BERT with Hybrid CNN do not come from chance variations but are statistically significant.

Table 2. Overall Performance Summary of Models

Model	Accuracy	Precision	Recall	Macro-F1	VAD RMSE
Baseline BERT	0.82	0.79	0.77	0.78	0.49
ABSA-LSTM	0.80	0.76	0.74	0.75	0.52
Hybrid CNN	0.85	0.82	0.80	0.81	0.44
Proposed Model	0.89	0.87	0.86	0.86	0.41

Table 2 Performance summary: Compared to all baseline models, the proposed Feature-Optimized BERT with Hybrid CNN achieves the highest accuracy, precision, recall, macro-F1, and the lowest VAD RMSE, proving that the incorporation of optimized embeddings and aspect-guided CNN features can greatly improve performance on both emotional classification and intensity prediction tasks. In general, the proposed model shows stronger generalization and becomes more reliable in understanding emotions in EmoBank.

Table 3. Comparative Study of Existing Methods vs. Proposed Model

Study / Model	Method Used	Emotion Task	Explainability Level	Limitations Compared to Present Work
(Devlin et al., 2019)	BERT (Transformer)	Generic emotion / sentiment	Low	No aspect-level analysis; limited token interpretability
(Wang et al., 2016)	ABSA-LSTM	Aspect-based sentiment	Medium	Weak contextual depth; no hybrid local-global modeling
(Kalyan et al., 2021)	Transformer Survey	Sentiment + emotion (review-level)	Conceptual only	No implementable framework; lacks explainability workflow
Hybrid CNN (Baseline)	Multi-kernel CNN	Emotion classification	Medium	Lacks global semantics from BERT embeddings
Present Work (2025)	Feature-Optimized BERT + Aspect-CNN	Multi-emotion + intensity + aspect	High (dual interpretability)	Integrates feature selection, aspect reasoning, and CNN saliency for superior performance

Table 3 following comparison table illustrates that BERT, ABSA-LSTM, and traditional CNNs are at best bound by one or more of the following limitations: contextual depth, aspect-awareness, and interpretability. This is in contrast to the proposed Feature-Optimized BERT with Aspect-Based Hybrid CNN, which offers richer emotional understanding, integrating global transformer semantics together with local aspect-sensitive cues. Finally, the comparison given above shows that the proposed model stands higher in accuracy, richness in explanations, and completeness of the framework while classifying and predicting the intensity of emotions.

V. MAJOR FINDINGS

It is clear that the proposed model significantly outperforms state-of-the-art baselines, including BERT, ABSA-LSTM, and hybrid CNNs, in accuracy and macro-F1 score, as well as VAD regression error. Feature optimization managed to shrink embedding dimensions by 40-60% without losing semantic richness and gained faster training with better generalization. More detailed evaluations of interpretability proved stronger alignment of emotional saliency and aspects-emotions mapping, confirming that design improvements boost both predictive performance and explainability.

VI. DISCUSSION

This improvement regarding emotional classification and intensity prediction originates mainly from the combination of semantically rich transformer embeddings with the local sensitivity of

CNN filters. Feature selection was important for removing noise from the high-dimensional embeddings and therefore helped the CNN to better detect the local emotional cues. This aspect extraction improved contextual anchoring of the emotional expressions with respect to target entities or topics and enhanced interpretability, hence allowing the model to outperform pure transformer architecture-based methods. The dual-explanation technique has further validated the fact that the model indeed attends to linguistically meaningful tokens and aspect spans, and this ensures trustworthiness and possible real-world usability. However, relying on dependency-based aspect extraction brings variability across domains, and for highly informal or multilingual text, it may require re-tuning.

VII. SCIENTIFIC CONTRIBUTIONS

The paper proposes an interpretable hybrid architecture that embeds optimized BERT embeddings into an aspect-aware CNN for enhancing fine-grained emotion detection and intensity modeling. The proposed approach handles categorical emotions and continuous VAD scores within one framework; thus, ensuring its versatility toward downstream tasks. Another contribution is the dual-level interpretability module that couples saliency analysis with Grad-CAM visualization, hence offering transparency in model reasoning. Finally, comprehensive statistical validation is presented, which shows improvements observed in performance are significant and reliable over evaluation folds.

VIII. CONCLUSION

These findings further establish that performance in emotion detection tasks is substantially improved by Feature-Optimized BERT along with an Aspect-Based Hybrid CNN, without compromising interpretability. Dimensionality reduction on the BERT embeddings boosts efficiency, while aspect-based CNN filters allow for finer emotional reasoning by capturing local sentiment cues bound to specific entities. Performance gains across various classification and intensity prediction tasks were statistically significant, with the interpretability tools provided verifying that the predictions were based on meaningful linguistic evidence. Overall, this framework strikes a very good balance between desired characteristics of accuracy, efficiency, and explainability.

IX. FUTURE WORK

Further work can extend the framework to include domain-adaptive pretraining that could enhance robustness across text in informal, multilingual, or low-resource domains. The approach can be further improved for aspect extraction through transformer-based sequence labelers or prompt-driven extraction with the aim of minimizing dependency parsing errors. Multitask learning could be explored to jointly predict emotions, aspects, and sentiment polarity and intensity for further enrichment of emotional comprehension. It can also be extended by including methods of contrastive learning or adapter-based fine-tuning with a view to reducing the cost of training while improving generalization. Finally, deployment into real-world applications tests user trust, the impact of interpretability, and system behavior in case of noisy or adversarial inputs.

REFERENCES

[1] Abas, A. R., Elhenawy, I., Zidan, M., & Othman, M. (2022). BERT-CNN: A Deep Learning Model for Detecting Emotions from Text. *Computers, Materials & Continua*, *71*(2). https://www.academia.edu/download/85738797/pdf.pdf

[2] Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. https://doi.org/10.1007/s10462-021-09958-2

[3] Aziz, K., Ji, D., Chakrabarti, P., Chakrabarti, T., Iqbal, M. S., & Abbasi, R. (2024). Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Scientific Reports*, *14*(1), 14646.

[4] Baziotis, C., Nikolaos, A., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S., & Potamianos, A. (2018). Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *Proceedings of The 12th International Workshop on Semantic Evaluation*, 245–255. https://aclanthology.org/S18-1037/

[5] Boer, K. de. (2024). *Towards Interpretable Multimodal Models for Emotion Recognition* [Master's Thesis].

https://studenttheses.uu.nl/handle/20.500.12932/46902

[6] Chutia, T., & Baruah, N. (2024). A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, *57*(8), 203. https://doi.org/10.1007/s10462-024-10831-1

[7] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://aclanthology.org/N19-1423/?utm_campaign=The+Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-_m9bbH_7ECE1h3lZ3D61TYg52rKpifVNjL4fvJ85uqggrXsWDBTB7YooFLJeNXHWqhvOyC

[8] *EmoBank*. (n.d.). Retrieved November 21, 2025, from https://www.kaggle.com/datasets/jackksoncsie/emobank?utm_source=chatgpt.com

[9] Fan, C., & Chen, H. (2025). Research on eXplainable Artificial Intelligence in the CNN-LSTM hybrid model for energy forecasting. *Journal of Building Engineering*, 113150.

[10] Friedrich, A., Xue, N., & Palmer, A. (2023). A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 599–622. https://aclanthology.org/2023.eacl-main.44/

[11] Frye, R. H., & Wilson, D. C. (2022). Comparative analysis of transformers to support fine-grained emotion detection in short-text data. *The International FLAIRS Conference Proceedings*, *35*. https://journals.flvc.org/FLAIRS/article/view/130612

[12] Geetha, M. P., & Renuka, D. K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks*, *2*, 64–69.

[13] Hayat, H., Ventura, C., & Lapedriza, A. (2019). On the use of interpretable CNN for personality trait recognition from audio. In *Artificial Intelligence Research and Development* (pp. 135–144). IOS Press. https://ebooks.iospress.nl/volumearticle/52829

[14] Hu, G., Lin, T.-E., Zhao, Y., Lu, G., Wu, Y., & Li, Y. (2022). UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv Preprint arXiv:2211.11256*. https://arxiv.org/abs/2211.11256

[15] Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). *AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing* (No. arXiv:2108.05542). arXiv. https://doi.org/10.48550/arXiv.2108.05542

[16] Kim, S., Shen, S., Thorsley, D., Gholami, A., Kwon, W., Hassoun, J., & Keutzer, K. (2022). Learned Token Pruning for Transformers. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 784–794. https://doi.org/10.1145/3534678.3539260

[17] Kim, Y. (2018). Convolutional neural networks for sentence classification. *arXiv Preprint arXiv:1408.5882*. https://arxiv.org/abs/1408.5882

[18] Lee, S., Han, D. K., & Ko, H. (2021). Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. *IEEE Access*, *9*, 94557–94572.

[19] Lv, Y., Wei, F., Zheng, Y., Wang, C., Wan, C., & Wang, C. (2021). A span-based model for aspect terms extraction and aspect sentiment classification. *Neural Computing and Applications*, *33*(8), 3769–3779. https://doi.org/10.1007/s00521-020-05221-x

[20] Makhmudov, F., Kultimuratov, A., & Cho, Y.-I. (2024). Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures. *Applied Sciences*, *14*(10), 4199.

[21] Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. *Proceedings of the 12th International Workshop on Semantic Evaluation*, 1–17. https://aclanthology.org/S18-1001/

[22] Shi, X., Yang, M., Hu, M., Ren, F., Kang, X., & Ding, W. (2025). Affective knowledge assisted bi-directional learning for Multi-modal Aspect-based Sentiment Analysis. *Computer Speech & Language*, *91*, 101755.

[23] Tellai, M., Gao, L., & Mao, Q. (2023). An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network. *International Journal of Speech Technology*, *26*(2), 541–557. https://doi.org/10.1007/s10772-023-10035-y

[24] Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 606–615. https://aclanthology.org/D16-1058.pdf

[25] Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., & Chen, Z. (2017). Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE Transactions on Knowledge and Data Engineering*, *18*(3), 320–333.

[26] Yuan, Y., Duo, S., Tong, X., & Wang, Y. (2025). A Multimodal Affective Interaction Architecture Integrating BERT-Based Semantic Understanding and VITS-Based Emotional Speech Synthesis. *Algorithms*, *18*(8), 513.

[27] Zhang, Y., Xu, H., Zhang, D., & Xu, R. (2024). A hybrid approach to dimensional aspect-based sentiment analysis using BERT and large language models. *Electronics*, *13*(18), 3724.