Context-Aware Sentiment Analysis of E-Commerce Reviews Using a BERT-CNN-BiLSTM Hybrid Model

Mahmud Reza Mahim, Fahad Bin Z Islam, Md Saklain Mahmud, Md. Imtiyaz Hasan, Sifat Rahman Ahona
Department of Computer Science and Engineering
American International University-Bangladesh (AIUB), Dhaka, Bangladesh

***Abstract*—User-generated product reviews are an essential source of information in e-commerce; nevertheless, the huge volume and varying quality of review texts make extracting insights difficult. The conventional approach to sentiment classification is limited in terms of recognizing contextual and aspect-oriented sentiment clues in the text. This study proposes a hybrid architecture that uses contextual sentiment clues for e-commerce reviews sentiment classification. The experiments are conducted using FABSA dataset which contains reviews annotated in terms of multiple aspects–sentiments pairs. Every review is expanded in terms of aspect-oriented sentiment classification samples. This makes it possible to learn fine-grained polarities of sentiments associated with specific aspects of product reviews. To tackle the problem of class imbalance in the data, the experiments employ the method of stratified oversampling in combination with the use of class-weighted cross-entropy loss function. The empirical results show that the improved hybrid BERT–CNN–BiLSTM model achieves 92.20% validation accuracy, 92.25% weighted F1-score, and 86.36% macro F1-score. The most noticeable progress has been made in the case of neutral class, where the F1-score has increased by 17.1 percentage points showing the improvements in minority-class recognition and decrease of class imbalance. The architecture-level ablation study shows the superior performance of the BERT–CNN–BiLSTM model compared to its simplified versions based on BERT, BERT–CNN, and BERT–BiLSTM architectures. A contextual comparison with reported FABSA baselines suggests competitive performance, although this comparison should not be interpreted as a strict leaderboard result because the baselines were not reproduced under identical experimental settings. Overall, the findings demonstrate that combining contextual transformer representations with local convolutional features and bidirectional sequential modeling can improve class-balanced sentiment classification in aspect-expanded e-commerce review data.**

***Keywords*—Sentiment analysis; e-commerce reviews; BERT; CNN; BiLSTM**

I. Introduction

A. Background and Motivation

The widespread adoption of e-commerce platforms has made user-generated product reviews one of the most influential sources of information in consumer purchase decisions [1], [2]. User-generated reviews help consumers reduce uncertainty regarding product quality, functionality, usability, and post-purchase satisfaction [2], [3]. Therefore, review information is increasingly used in recommendation and decision-support systems. However, the abundance of review content creates information overload, making it difficult for consumers and online platforms to filter useful information from a large pool of reviews. In addition, many reviews are short, generic, sentiment-heavy, or low in informational value, while detailed experience-based reviews may remain unnoticed [4], [5], [6], [7].

Traditional sentiment analysis methods generally focus on assigning broad polarity labels, such as positive, negative, or neutral, to a given text. Although such approaches are useful, they often fail to capture fine-grained contextual information related to specific product aspects, usage conditions, and user intentions [8], [9]. For example, a review may express positive sentiment toward one aspect of a product while expressing negative sentiment toward another. Similarly, short expressions, such as "Great product!", may indicate positive polarity but provide limited contextual detail about the product experience. These cases show that sentiment classification in e-commerce reviews requires models that can capture not only overall polarity, but also the contextual cues surrounding product-related aspects [6], [7], [10], [11].

Lexicon-based and conventional machine learning methods have provided useful and interpretable baselines for sentiment analysis. Some recent studies also show that calibrated lexicon-based approaches can remain competitive in specific domains [9]. However, conventional models often struggle with domain-specific language, compositional meaning, sarcasm, implicit sentiment, and the semantic relationship between product aspects and opinion expressions [8], [10]. These limitations have encouraged the use of deep learning and transformer-based methods, which are more capable of learning contextual representations directly from text.

Recent studies have increasingly adopted deep learning architectures to capture both local and global dependencies in review text. BERT provides contextualized word representations through bidirectional transformer encoding, making it effective for modeling semantic relationships in sentences [12]. Convolutional neural networks (CNNs) are effective at extracting local n-gram and phrase-level sentiment features, while bidirectional long short-term memory networks (BiLSTMs) can capture sequential dependencies from both forward and backward directions [8], [13], [3]. Hybrid architectures that combine these components can integrate contextual representation learning, local feature extraction, and sequential dependency modeling within a single framework [13], [14], [15], [16].

In this context, a hybrid BERT–CNN–BiLSTM architec-

ture offers a promising approach for context-aware sentiment classification in e-commerce reviews. BERT can generate deep contextual embeddings, CNN layers can identify local sentiment-bearing patterns, and BiLSTM layers can model bidirectional dependencies across review sequences. This combination is particularly suitable for aspect-expanded review data, where each review may contain multiple aspect–sentiment pairs. Therefore, this study focuses on developing and evaluating a BERT–CNN–BiLSTM hybrid model for sentiment classification using the FABSA dataset, with emphasis on validation performance, class-balanced evaluation, ablation analysis, and cross-dataset generalization.

B. Literature Review

1) From lexicon-based to deep learning approaches: Sentiment analysis has evolved from lexicon-based techniques and classical machine learning to deep learning approaches. Early systems relied on handcrafted sentiment lexicons and surface-level representations such as bag-of-words, n-grams, and shallow syntactic cues [8], [9]. Although these methods offered a degree of interpretability and domain independence, they struggled with domain adaptation, context-dependent polarity, and subtle semantic phenomena such as irony and implicit sentiment [8], [9], [10].

Recent studies, such as MultiLexScaled, show that aggregated and rescaled lexicon-based methods can still yield competitive and interpretable sentiment scores for large review datasets. However, these methods struggle to fully understand complex compositional statistics and non-additive interactions among sentiment indicators and product-specific attributes in single reviews [9], [10].

Advances in the field of deep learning have strengthened sentiment analysis in terms of end-to-end representation learning from raw text. Under the review dataset benchmark [8], CNN-based models automatically extract local sentiment-bearing patterns as well as phrase-level features. Additionally, recurrent models like LSTM and BiLSTM have greater capacity than traditional methods to capture long-distance dependencies and sequential dynamics. Different hybrid architectures, such as CNN + LSTM and BiLSTM with attention, make sentiment analysis more robust by leveraging the synergy between complementary feature extractors [8], [17]. Early work consistently demonstrated that hybrid deep learning models outperformed single baseline models across tasks requiring understanding of text, images, and multimodal data for sentiment analysis [8], [10].

2) Transformer and hybrid architectures for sentiment analysis: The rise of pre-trained transformers, particularly BERT, was a turning point for sentiment and emotion detection, as they enable expressive contextual representations through self-attention and large scale pre-training [10], [12]. BERT based fine-tuning has achieved strong results across product review analysis, political debate analysis, and social media monitoring, often with limited task-specific data [1], [11], [18]. More recent extensions, such as RoBERTa-BiLSTM and ontology-enhanced BERT, show that combining transformer embeddings with sequential modeling or external knowledge can yield additional gains, particularly in domain-specific and aspect-level settings [10], [11], [13]. Survey evidence also suggests that transformer based and hybrid approaches now dominate leading results, with F1-scores often exceeding 0.85 on well-annotated datasets [10].

Recent research has increasingly combined BERT with CNN and BiLSTM architectures to leverage their complementary inductive biases, with CNNs capturing local n-gram features and BiLSTMs modeling sequential dependencies. Liu *et al.* proposed a BERT–BiGRU–Softmax model for e-commerce product reviews and showed that contextual embeddings combined with gated recurrent units can outperform strong recurrent baselines [1]. Shen introduced a BERT–CNN–BiLSTM–Attention model for sentiment analysis of social media texts related to aquatic product companies and reported an F1-score above 93% [14]. Khan *et al.* combined multilingual BERT with stacked CNN–BiLSTM and attention for Urdu sentiment analysis, demonstrating the usefulness of hybrid architectures in low-resource and morphologically complex settings [15]. Using this architectural pattern, they utilized a CNN–BiLSTM–BERT hybrid model and analyzed multilingual university feedback captured through Twitter, reporting above 92% accuracy. Their results demonstrate that this architectural pattern is scalable across different domains as well as languages [16]. This approach applies to sentiment and emotion detection on handwritten text, as well as text in electronic format, suggesting that these hybrid models are highly versatile [17].

Overall, the literature shows that hybrid architectures composed of pretrained transformers layered with convolutional layers and recurrent components are more effective than their single-architecture baselines in terms of accuracy and general robustness across heterogeneous sentiment-analysis benchmarks [8], [14], [15], [13], [16].

3) Aspect-level and context-aware sentiment modeling: In e-commerce review analysis, sentiment classification becomes more informative when it moves beyond review-level polarity and considers the specific product aspects discussed in the text. A single review may contain different sentiments toward different aspects of the same product, such as positive sentiment toward battery life but negative sentiment toward price or durability. Therefore, aspect-level and context-aware sentiment modeling is important for capturing fine-grained opinion patterns in product reviews [6], [7], [19].

The FABSA dataset provides a suitable benchmark for this task because it represents user reviews through aspect–sentiment annotations, allowing each review to be analyzed in relation to the product aspects it contains [19]. This type of data is especially useful for evaluating models that need to learn sentiment polarity from contextual cues rather than relying only on surface-level sentiment words. However, aspect-expanded sentiment classification also introduces several challenges, including class imbalance, contextual ambiguity, and difficulty in recognizing minority classes such as neutral sentiment [6], [20].

Previous studies have shown that transformer-based models and hybrid deep learning architectures can improve sentiment classification by learning richer representations from review text [1], [2], [3]. BERT-based models are effective for capturing contextual semantics, while CNN and BiLSTM components can further support local phrase-level feature extraction and

bidirectional sequence modeling [14], [15], [13]. Nevertheless, the combined use of BERT, CNN, and BiLSTM for aspect-expanded sentiment classification in e-commerce reviews remains an important direction for further investigation.

This gap motivates the present study, which develops and evaluates a hybrid BERT-CNN-BiLSTM model for context-aware sentiment classification on the FABSA dataset. The study focuses on validation performance, class-balanced evaluation, architecture-level ablation, and cross-dataset generalization. By doing so, it aligns the implemented experimental scope with the stated objective of improving sentiment classification in aspect-expanded e-commerce review data.

C. Research Aims and Contributions

The primary aim of this study is to design and evaluate a hybrid BERT–CNN–BiLSTM model for context-aware sentiment analysis and demonstrate that the hybrid formulation is suitable for aspect-expanded review data. More specifically, the study investigates whether combining transformer-based contextual encoding with convolutional and bidirectional sequential modeling can improve sentiment classification on aspect-expanded review instances derived from the FABSA dataset [1], [2], [14], [15], [13], [19]. In the proposed architecture, BERT provides contextual embeddings, CNN layers capture local phrase-level features, and BiLSTM layers model sequential dependencies across the review text [1], [14], [15]. This combination is intended to capture both localized sentiment cues and broader contextual structure within user reviews.

1) Research objectives:

O1: **Hybrid Model Design:** Develop a BERT–CNN–BiLSTM architecture for sentiment classification on aspect-expanded e-commerce review data, using BERT for contextual encoding, CNN for local feature extraction, and BiLSTM for bidirectional sequence modeling [1], [14], [15], [13].

O2: **Validation-Based Performance Evaluation:** Evaluate the proposed model on the FABSA dataset using standard classification metrics, including accuracy, weighted F1- score, macro F1-score, class-wise precision, recall, F1- score, and confusion-matrix analysis [19], [6], [20].

O3: **Comparative Performance Analysis:** Compare the optimized hybrid configuration with the baseline training configuration and contextualize its performance against reported results from existing sentiment-analysis models in the literature [1], [2], [3], [14].

2) Research contributions:

C1: **Hybrid BERT–CNN–BiLSTM Architecture for Sentiment Classification.** This study develops and implements a hybrid BERT–CNN–BiLSTM architecture for context-aware sentiment classification in e-commerce reviews. In the proposed model, BERT is used for contextual representation learning, CNN layers capture local sentiment-bearing phrase patterns, and BiLSTM layers model bidirectional sequential dependencies [1], [14], [15].

C2: **Aspect-Expanded Evaluation on the FABSA Dataset.** The study evaluates the proposed model on the FABSA dataset by converting review-level aspect–sentiment annotations into aspect-expanded sentiment classification instances. This allows the model to be assessed on fine-grained sentiment labels associated with product-review aspects [19].

C3: **Empirical Performance Analysis Using Class-Balanced Metrics.** The proposed model is evaluated using accuracy, weighted F1-score, macro F1-score, class-wise precision, recall, F1-score, and confusion-matrix analysis. The optimized configuration improves macro F1-score and minority-class recognition, particularly for the neutral class, indicating more balanced sentiment classification performance [6], [20].

C4: **Architecture-Level Ablation Study.** An ablation study is conducted to compare BERT, BERT–CNN, BERT–BiLSTM, and the full BERT–CNN–BiLSTM architecture under the same experimental setting. The results show that the full hybrid model achieves the strongest performance, demonstrating the complementary contribution of convolutional and recurrent components.

C5: **Cross-Dataset Generalization Analysis.** The study further examines the generalization behavior of the trained model through zero-shot evaluation on the SemEval-2014 Restaurant dataset. This analysis highlights the model's ability to transfer polarity recognition across domains while also revealing the domain-sensitive nature of neutral sentiment classification.

3) Research questions: In line with the implemented scope of the study, the following research questions guide the empirical investigation:

RQ1: How effectively can a hybrid BERT–CNN–BiLSTM model classify sentiment in aspect-expanded e-commerce reviews [1], [14], [15], [13]?

RQ2: To what extent does optimizing the hybrid training configuration improve overall and class-balanced sentiment performance compared with the initial baseline configuration [6], [20]?

RQ3: How does the proposed hybrid model compare, in contextual terms, with reported results from existing machine learning, deep learning, and transformer-based sentiment-analysis models on the FABSA benchmark [19], [1], [2], [3]?

By addressing these questions, the study seeks to improve understanding of how hybrid transformer–CNN–BiLSTM architectures can be applied to sentiment classification in e-commerce review analysis. [1], [2], [6], [14], [15], [21].

The remainder of this study is organized as follows: Section II presents the dataset, preprocessing steps, proposed BERT–CNN–BiLSTM architecture, training configuration, and evaluation metrics. Section III reports the experimental results, including validation performance, ablation analysis, class-wise evaluation, contextual comparison with reported FABSA

baselines, and cross-dataset validation. Section IV discusses the findings and interprets the strengths and limitations of the proposed model. Section V presents the limitations of the study. Finally, Section VI and Section VII provide the conclusion and future research directions, respectively.

II. Methodology

A. Dataset and Data-Preprocessing

1) Dataset: Experiments utilize the FABSA (Fine-grained Aspect-Based Sentiment Analysis) dataset, obtained from the Hugging Face Datasets Hub (`jordiclive/FABSA`). This dataset comprises product review texts accompanied by structured labels containing aspect–sentiment pairs, where each label specifies a product aspect (such as *battery*, *screen*, or *price*) and its corresponding sentiment polarity (*positive*, *negative*, or *neutral*). The dataset is pre-divided into training, validation, and test splits, which are maintained throughout all experiments to ensure fair and reproducible evaluation.

2) Data pre-processing pipeline: The raw dataset is subjected to a multi-stage pre-processing pipeline prior to input into the hybrid model.

a) Aspect–sentiment extraction: Every instance in the FABSA corpus consists of free-form reviews and labels containing pairs of aspects and sentiments. In the process of pre-processing, the reviews and their respective aspect and sentiment pairs are extracted from all the samples in the corpus. The result is a structured format with the review text in each instance and (aspect, sentiment) pairs for that review.

b) Class imbalance analysis: Before resampling, the frequency of each sentiment class (*positive*, *negative*, *neutral*) was calculated for the entire dataset, which consisted of the training, validation, and test data. The outcomes highlight a significant imbalance in class representation, with a disproportionately small number of neutral samples compared to the dominant positive class.

c) Class imbalance correction via upsampling: Class imbalance became evident in the training dataset. To address this issue, the dataset was divided into three sentiment-based groups. Stratified random sampling with replacement was then used to up-sample the minority neutral group, increasing its number to be almost equal to the number of items in the majority positive group. No changes were made to the negative group.

d) Dataset expansion: The reviews can not be mapped directly to their respective labels (which include the target words), since one review may contain more than one aspect-sentiment pairs; hence, multi-label learning is necessary. The (review, aspect, sentiment) tuple is considered as one instance, and the generation of new instances comes from the tuples to add another layer of granularity in the learning task. In particular, for each aspect-based review having k number of aspects, there will be k copies (one for each aspect) together with its respective sentiment label.

e) 80–20 Split alignment: After aspect-based expansion, the size of the augmented training set became disproportionately larger than the expanded test set. To maintain an approximate 80:20 train-to-test ratio, a uniform subsample was drawn from the augmented training set without replacement, making the final training set four times larger than the expanded test set. The validation set was kept unchanged throughout this process.

f) Tokenization: Tokenization for all review texts is performed through the BERT-base (uncased) WordPiece tokenizer. Every sequence out of the three is padded to ensure that they satisfy the condition of having a maximum sequence length as per the batch and truncated to 512 tokens, which is the total number of positional embeddings that characterize the BERT architecture. The tokenizer produces token indices and attention masks, where the attention mask distinguishes actual tokens from padding tokens.

g) Label encoding: The three sentiment polarities are mapped to integer indices: *negative* $\rightarrow$ 0, *neutral* $\rightarrow$ 1, and *positive* $\rightarrow$ 2. This encoding is applied consistently across all splits and is utilized by the cross-entropy loss function during training.

h) Class weight computation: In order to reduce the imbalance in the dataset even after the use of upsampling, class weights are calculated using the inverse frequency formula:

$$w_c = \frac{N}{N_c} \tag{1}$$

where, N denotes the total number of training instances and N_c represents the number of instances belonging to class c. Hence, in this case, the assigned weights are 1.5 (negative), 2.0 (neutral), and 1.0 (positive).

i) Dataset and data loader construction: The tokenized sequence and its corresponding label in integer format are arranged in a structured dataset, which provides individual tensor samples upon request. The training, validation, and testing sets each had their own data iterator created using a batch size of 8. For the training data iterator, randomness is applied to shuffle the data for every epoch to avoid any bias due to order, while the validation and testing data iterators follow a sequential flow.

B. Base Models

1) BERT: Bidirectional Encoder Representations from Transformers (BERT) refers to a pre-trained language model that considers both left and right-contextual information in interpreting text data [12]. The model has a multi-layered bidirectional Transformer encoder, which can be of different sizes, like $\text{BERT}_{\text{BASE}}$ (12 layers, 768 hidden units, 12 self-attention heads, 110 million parameters) and $\text{BERT}_{\text{LARGE}}$ (24 layers, 1024 hidden units, 16 self-attention heads, 340 million parameters). The inputs are fed into the network using WordPiece tokenization, where each input consists of a vocabulary of 30,000 words and includes special tokens like [CLS], used for classification tasks, and [SEP], used to differentiate sentences. Token, segment, and positional embeddings are used in the model. The pre-training phase incorporates two strategies. Masked Language Modeling (MLM) hides 15% of tokens and tasks the model with predicting them. Meanwhile, Next Sentence Prediction (NSP) helps the model to recognize sentence interactions. Fine-tuning involves task-specific training and parameter tuning. Fig. 1 presents the BERT model architecture.

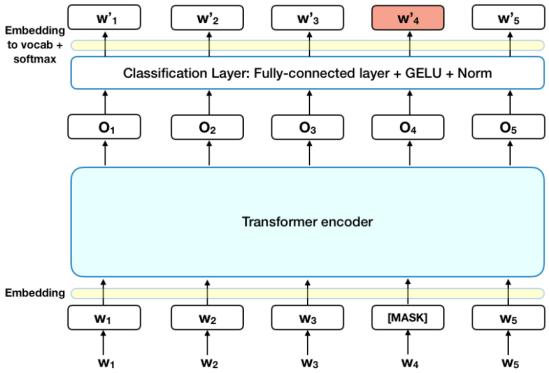


Fig. 1. BERT model architecture.

2) CNN: The process of text classification using CNNs involves the application of convolutions on the vectors of words in the sentence [22]. In this regard, each sentence is viewed as a matrix of word vectors where each word can be defined as a k-dimensional vector, hence an n x k matrix. Various filters of different sizes, for example, 3, 4, or 5, are used to produce feature maps by sliding over the matrix. In doing so, the filter uses activation functions such as ReLU or tanh. The next step involves performing the max-over-time pooling function on the obtained feature maps in order to select the most important features while at the same time allowing for the variation in sentence length. All the extracted features from various filters are concatenated to obtain a vector that is used as input to the fully connected softmax layer for classification purposes. Some models may use several channels with predetermined and trainable embeddings (see Fig. 2).

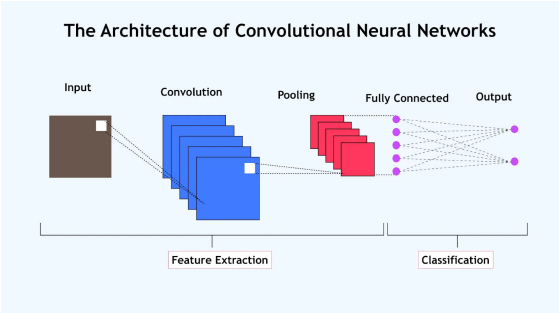


Fig. 2. CNN architecture for text classification.

3) Bi-LSTM: The Bidirectional Long Short Term Memory (Bi-LSTM) network refers to an advancement over the traditional LSTM model by processing the sequence in the forward and backward direction, thereby enabling context from past as well as future tokens [23]. It consists of two LSTM layers, where one layer processes the sequence from left to right and the other one processes it from right to left. Each LSTM unit includes three gates (input, forget, output) that regulate information retention over time. At every time step, the network concatenates the outputs from both layers. In Natural Language Processing (NLP), the Bi-LSTM models can be used for applications like sequence tagging, sentiment analysis, and text classification since they handle dependencies in both directions (see Fig. 3).

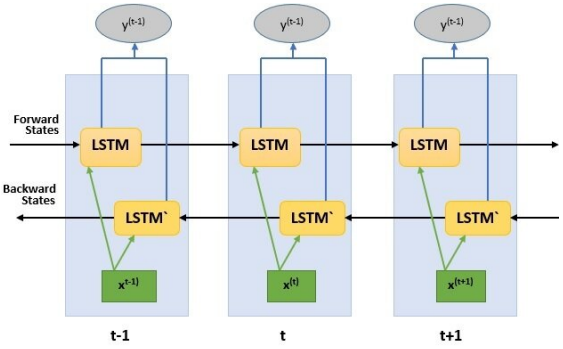


Fig. 3. Bi-LSTM architecture.

C. Hybrid Model Architecture

The proposed hybrid model consists of BERT to generate contextual embeddings, a CNN to extract local features, and a Bi-directional LSTM (Bi-LSTM) to capture sequences. These components are used for conducting context-based sentiment analysis of reviews. The generated BERT representations are passed through two modeling stages. First, three parallel CNN branches with kernel sizes of 3, 4, and 5 extract local n-gram sentiment features from the contextual token sequence. The CNN outputs are passed through ReLU activation and layer normalization before being concatenated along the feature dimension. The concatenated representation is then passed to a bidirectional LSTM to model sequential dependencies in both forward and backward directions. Finally, global max-pooling is applied over the sequence dimension, followed by dropout and a linear classification layer for three-class sentiment prediction (see Fig. 4).

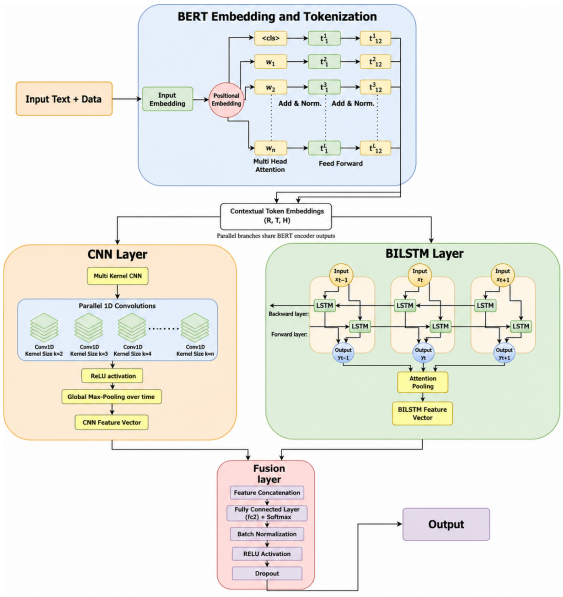


Fig. 4. Hybrid BERT-CNN-BiLSTM model architecture.

The following hyperparameters define the final implemented architecture. The CNN branch consists of three parallel convolutional layers (`Conv1d`) applied to the BERT last-hidden-state sequence (`bert-base-uncased`, hidden

dimension 768), each producing 256 output feature maps (`out_channels` $= 256$) with kernel sizes of 3, 4, and 5 tokens respectively, with padding values of 1, 2, and 2 to maintain compatible output lengths. Each convolutional output is passed through a ReLU activation followed by a `LayerNorm` layer before being fed into the shared BiLSTM. The three branch outputs—each of shape [batch $\times$ seq_len $\times$ 256]—are concatenated along the feature axis, giving a BiLSTM input size of 768 ($= 256 \times 3$). The BiLSTM has a hidden size of 128 units per direction (`bidirectional = True`), yielding a 256-dimensional output per token. Global max-pooling is then applied over the sequence dimension, followed by a dropout layer ($p = 0.5$) and a linear classification head projecting to three sentiment classes. Table I summarises these values.

TABLE I. HYPERPARAMETER SUMMARY OF THE BERT–CNN–BILSTM HYBRID MODEL.

Parameter	Value
BERT encoder	`bert-base-uncased`
BERT hidden dimension	768
Number of CNN branches	3
CNN output channels	256 per branch
CNN kernel sizes	3, 4, and 5
CNN padding	1, 2, and 2
CNN activation	ReLU
Normalization	LayerNorm after each CNN branch
Concatenated CNN dimension	768 (256×3)
BiLSTM input size	768
BiLSTM hidden units	128 per direction
BiLSTM direction	Bidirectional
BiLSTM output dimension	256
BiLSTM layers	1
Pooling method	Global max-pooling
Dropout rate	0.5
Classifier output	Linear layer ($256 \rightarrow 3$)
Batch size	8
Learning rate	1×10^{-5}
Training epochs	20
BERT fine-tuning	Frozen for first 5 epochs
Loss function	Weighted cross-entropy
Class weights	Negative: 1.5, Neutral: 2.0, Positive: 1.0

D. Model Evaluation Metrics

The performance of the proposed hybrid model is evaluated using standard classification metrics for multi-class sentiment analysis. Since the task involves three sentiment classes, namely negative, neutral, and positive, the evaluation considers both overall performance and class-wise effectiveness. The main metrics used in this study are Accuracy, Precision, Recall, F1-score, Macro-F1, Weighted-F1, and confusion-matrix analysis [20].

Accuracy measures the proportion of correctly classified samples among all samples:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2}$$

For each sentiment class, Precision measures the proportion of correctly predicted samples among all samples predicted as that class. Recall measures the proportion of correctly predicted samples among all actual samples of that class. The F1-score represents the harmonic mean of Precision and Recall. These metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. In the multi-class setting, these values are calculated for each sentiment class using a one-versus-rest interpretation.

Because sentiment datasets are often imbalanced, especially when one class has fewer samples than the others, macro-averaged and weighted-averaged metrics are also used. Macro-averaged metrics assign equal importance to each class by taking the simple average of the class-wise scores:

$$\text{Macro-Precision} = \frac{1}{C} \sum_{i=1}^{C} \text{Precision}_i \tag{6}$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{i=1}^{C} \text{Recall}_i \tag{7}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{8}$$

where, C denotes the total number of sentiment classes.

Weighted-F1 is also reported to account for class distribution by weighting each class-specific F1-score according to its support:

$$\text{Weighted-F1} = \sum_{i=1}^{C} \frac{n_i}{N} \times \text{F1}_i \tag{9}$$

where, n_i is the number of samples in class i, and N is the total number of samples.

In addition to these numerical metrics, confusion matrices are used to analyze class-wise prediction behavior. The confusion matrix helps identify which sentiment classes are correctly classified and which classes are more frequently confused with one another. This is particularly important for evaluating the model's performance on minority classes such as neutral sentiment.

III. RESULTS AND ANALYSIS

In this section, the empirical results are discussed, based on the findings in the experimental setup related to the proposed BERT-CNN-BiLSTM hybrid model. In the implemented experiment, each review is expanded into aspect-text pairs, and the hybrid model is trained based on this aspect-wise partitioning approach. The results can thus be seen as part of the sentiment classification portion of this study.

A. Overall Validation Performance

Two configurations were examined with a normal hybrid model and the optimized hybrid model. As shown in Table II, optimization improved all headline validation metrics. Validation accuracy increased from 88.75% to 92.20%, weighted F1 increased from 89.32% to 92.25%, and macro F1 increased from 78.64% to 86.36% (see Fig. 5). The improvement in macro F1 is especially significant, because it indicates that the opti- mized configuration improved class balance and minority-class recognition rather than only strengthening performance on the dominant class. in the optimized configuration it dropped to 5.89 percentage points. This narrowing suggests that the optimized model produced more balanced decisions across the three sentiment classes.

TABLE II. VALIDATION PERFORMANCE OF THE ORIGINAL AND OPTIMIZED HYBRID MODELS

Configuration	Accuracy	Weighted F1	Macro F1
Original hybrid model	0.8875	0.8932	0.7864
Optimized hybrid model	0.9220	0.9225	0.8636

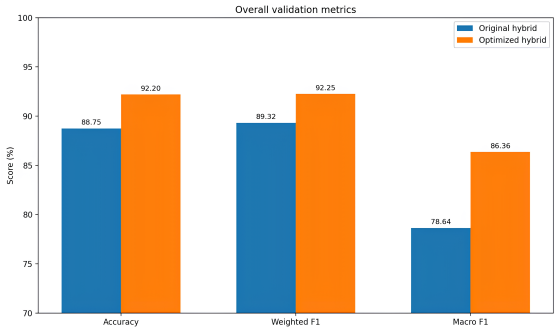


Fig. 5. Bar-chart comparison of overall validation metrics for the original and optimized hybrid models.

One of the explanations for such an increase is to examine the difference between the weighted F1 and macro F1 scores. Specifically, in the initial setting, the value of this difference was 10.68%, while after optimization, it dropped down to 5.89%. It shows that in the optimized setting, decisions on each class were more evenly distributed.

B. Architecture-Level Ablation Experiments

In order to make sure that the final performance gain achieved is due to the full hybrid architecture and not just BERT, an ablation study has been performed. Four models have been compared with the help of the exact same preprocessing approach and the exact same held out FABSA testing dataset, namely: 1) BERT, 2) BERT+CNN, 3) BERT+BiLSTM, and 4) BERT-CNN-BiLSTM. For all these tests, F1-score means weighted average test F1 from the final comparison.

According to the ablation experiments' findings, the best F1 value for testing data belongs to the complete BERT-CNN-BiLSTM architecture, whose F1 is 0.9286. On the other hand, BERT+CNN and BERT+BiLSTM models give 0.9196 and 0.9182 F1 values, respectively, while using only BERT gives an F1 value of 0.9167. Therefore, the hybrid model improved over the strongest reduced variant, BERT+CNN, by approximately 0.0090 absolute F1. The improvement is modest but consistent, suggesting that the convolutional and recurrent branches contribute complementary information as CNN layers capture local sentiment-bearing n-gram features, while the BiLSTM branch models bidirectional sequential dependencies across the review text (see Table III).

TABLE III. ABLATION EXPERIMENT RESULTS ON THE FABSA TEST SET

Model	F1-score
BERT	0.9167
BERT+CNN	0.9196
BERT+BiLSTM	0.9182
BERT-CNN-BiLSTM model	0.9286

These findings confirm the proposed architecture. While BERT provides strong context embeddings, the best F1 value is achieved only if BERT vectors are used together with the CNN and BiLSTM modules. Hence, the results support the effectiveness of the proposed architecture.

C. Learning Behavior and Convergence

In comparison, the improved architecture is able to outperform the initial configuration significantly. While the base hybrid neural network with hyperparameter tuning continues to improve in performance throughout the 20 epochs, albeit more slowly, to yield a weighted F1 score of 79.80%, the optimized neural network starts off on an advantageous note and scores over 90% validation weighted F1 from epoch 3 onwards (as shown in Fig. 6), with its peak being reached at epoch 10 at 92.68% accuracy, 92.76% weighted F1, and 0.2423 validation loss. From epoch 10 onward, the validation loss slightly increases compared to previous epochs, whereas training accuracy keeps improving, signaling moderate overfitting towards the end of the epochs. Considering the separation variable analysis approach, it can be concluded that early stopping at epoch 10 would have been optimal to retain maximal generalization efficiency while cutting down on training expenses.

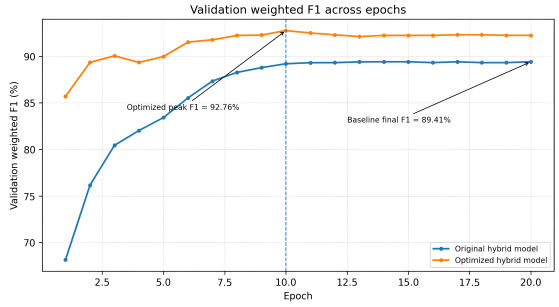


Fig. 6. Validation weighted F1 across epochs for the original and optimized hybrid models.

D. Contextual Comparison with Reported FABSA Baselines

In order to contextualize the proposed model, Table IV compares its performance with baseline and transformer-based

models reported on the FABSA benchmark by Kontonatsios et al. [19]. The proposed BERT–CNN–BiLSTM model achieves precision, recall, and F1-score values of 0.840, 0.890, and 0.870, respectively, which indicate competitive performance in this contextual comparison.

TABLE IV. CONTEXTUAL COMPARISON WITH REPORTED FABSA MODELS.

Model	Precision	Recall	F1-score
LogReg-BoW	0.784	0.491	0.604
GRU-GNN	0.702	0.667	0.684
BERT-single-base	0.785	0.745	0.765
BERT-single-large	0.785	0.792	0.788
BERT-PT	0.785	0.792	0.788
GAS	0.785	0.779	0.782
RoBERTa-single-base	0.781	0.787	0.784
RoBERTa-single-large	0.792	0.816	0.804
RoBERTa-pair-base	0.796	0.764	0.779
RoBERTa-pair-large	0.807	0.792	0.800
DeBERTa-single-base	0.777	0.787	0.782
DeBERTa-single-large	0.791	0.820	0.805
DeBERTa-pair-base	0.793	0.804	0.798
DeBERTa-pair-large	0.806	0.812	0.809
BERT-CNN-BiLSTM Hybrid Model	**0.840**	**0.890**	**0.870**

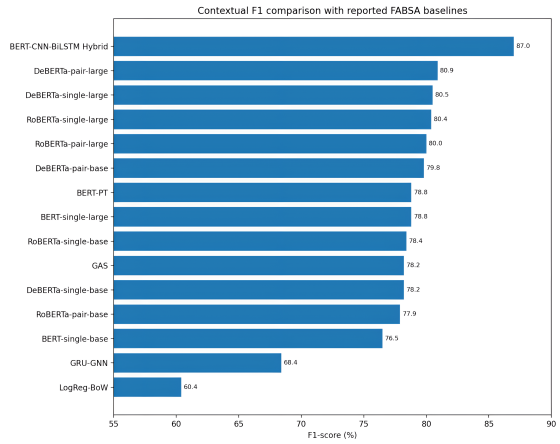


Fig. 7. F1-score comparison between the proposed model and reported FABSA benchmark models.

The proposed model obtains an F1-score of 0.870 in this contextual comparison, while the highest reported baseline F1-score in Table IV is 0.809 for DeBERTa-pair-large. This suggests that the hybrid architecture is promising for aspect-expanded sentiment classification. However, since the baseline models were not reimplemented under the same preprocessing, training, and evaluation conditions, this result should not be interpreted as a definitive ranking. A strict leaderboard comparison would require all models to be evaluated under identical experimental settings (see Fig. 7).

E. Class-Wise Validation Analysis

The optimized model proves to be most effective for the positive class, exhibits consistency for the negative class, and reaches its highest percentage gain for the neutral class. The values of precision, recall, F1-score, and support of the optimized model on the validation set are shown in Table V.

TABLE V. CLASS-WISE VALIDATION PERFORMANCE OF THE OPTIMIZED HYBRID MODEL

Class	Precision	Recall	F1-score	Support
Negative	0.8863	0.8675	0.8768	566
Neutral	0.6914	0.8485	0.7619	66
Positive	0.9534	0.9511	0.9522	1226

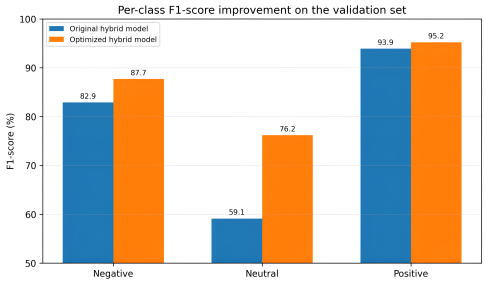
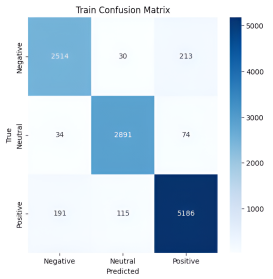
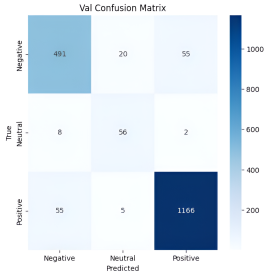


Fig. 8. Per-class F1-score comparison between the original and optimized hybrid models.

As seen in Fig. 8, the highest improvement is observed for the neutral category, with its F1 score growing from 59.1% to 76.2%, which means the absolute improvement reaches 17.1%. For the negative class, the F1 score goes up from 82.9% to 87.7%, and for the positive one, the F1 score jumps from 93.9% to 95.2%. Therefore, this finding confirms the conclusions that can be drawn based on the macro F1 score results.



(a) Training confusion matrix



(b) Validation confusion matrix

Fig. 9. Training and validation confusion matrices of the optimized BERT-CNN-BiLSTM model.

The confusion matrices provided in Fig. 9 show that the optimized model achieves very good class separation on both the train and validation splits, with observations forming a

cluster along the main diagonal of the confusion matrix. On the train data, the number of correctly classified examples for the negative, neutral, and positive classes is 2,514, 2,891, and 5,186 respectively, which shows that the model can learn discriminative representations for all three classes. On the validation split, there are 491 correct predictions for the negative class out of 566, 56 for the neutral class out of 66, and 1,166 for the positive class out of 1,226. The main misclassification sources occur in the classification of positive and negative sentiments, which results in 55 examples being misclassified from one of these two classes to another. While the neutral class is predicted quite accurately, a small portion of negative and positive examples is incorrectly classified as belonging to the neutral class.

F. Cross-Dataset Validation

Cross-dataset validation was used to evaluate generalization. Optimized BERT-CNN-BiLSTM model trained using FABSA dataset was tested on SemEval-2014 Restaurant data without any further training on the external data. SemEval labels were converted for sentiment classification into the same three-class label set used in FABSA dataset: negative, neutral, and positive classes. The SemEval-2014 Restaurant split used in this evaluation consists of 3,602 instances, of which 2,164 were positive, 805 were negative, and 633 were neutral (see Table VI).

TABLE VI. CROSS-DATASET VALIDATION SUMMARY

Evaluation data	Setting	Samples	Accuracy	Weighted F1	Macro F1
FABSA test set	In-domain held-out test	2,812	0.9282	0.9286	0.88
SemEval-2014 Restaurant	Zero-shot cross-dataset	3,602	0.7160	0.6547	0.51

The SemEval-2014 zero-shot evaluation achieves an accuracy of 0.7160 and weighted F1 score of 0.6547. Comparing this to the in-domain FABSA weighted F1 score of 0.9286, the external testing achieves an absolute reduction in F1 score by about 0.2740. This can be expected because of the existence of the domain shift problem where the model has been trained using FABSA dataset reviews but tested using SemEval review without being externally fine-tuned. Despite this domain shift, there still remains considerable sentiment transfer abilities in detecting polarity sentiment.

TABLE VII. CLASS-WISE SEMEVAL-2014 ZERO-SHOT VALIDATION REPORT.

Class	Precision	Recall	F1-score	Support
Negative	0.58	0.73	0.65	805
Neutral	0.42	0.02	0.04	633
Positive	0.77	0.91	0.84	2,164

The class-wise results obtained on SemEval-2014 show that transferability is highest for the positive class (F1 $= 0.84$) and moderate for the negative class (F1 $= 0.65$). However, the neutral class exhibits a near-total performance collapse, with an F1-score of only 0.04, which requires further analysis. This result can be explained by annotation-level semantic divergence between the two datasets. In FABSA, neutral sentiment is defined at the aspect level, where a reviewer may be neutral toward a specific product feature while expressing strong sentiment toward other aspects. In SemEval-2014, neutral polarity is often assigned at the sentence level and may represent factual or descriptive statements without explicit sentiment expression. As a result, the model trained on FABSA-style neutral instances encounters a qualitatively different neutral category during zero-shot transfer. The precision–recall pattern in Table VII further supports this interpretation: neutral precision is 0.42, but recall is only 0.02, indicating that the model rarely assigns the neutral label. Although the class-weighting scheme used during training was intended to improve neutral-class recall within the FABSA domain, the model appears to default to polar predictions when exposed to the more domain-specific vocabulary and sentence-level annotations of SemEval-2014.

A class-prior mismatch may also contribute to this behavior. Neutral instances represent only 17.6% of the SemEval-2014 Restaurant test split, with 633 neutral samples out of 3,602 total instances, whereas the FABSA training distribution was altered through upsampling to improve minority-class representation. This distributional difference may reinforce the model's tendency to avoid neutral predictions under external-domain inference. Taken together, annotation divergence, learned polar decision bias, and class-prior mismatch explain the observed neutral F1-score of 0.04. This finding does not invalidate the overall cross-dataset result, as the model still achieves 71.6% accuracy and a weighted F1-score of 0.65, but it shows that neutral sentiment representations are highly domain-sensitive. Future work should address this issue through aspect-conditioned input formatting, where the aspect term is explicitly included during tokenization, and through domain-adaptive fine-tuning on target-domain neutral examples to recalibrate the decision boundary for this class.

G. Result Interpretation

The experimental results indicate that the proposed hybrid configuration based on the combination of BERT, CNN, and BiLSTM architectures is effective for sentiment classification in the aspect-expanded FABSA review data set. The optimized configuration provided 0.9220 validation accuracy, 0.9225 weighted F1-score, and 0.8636 macro F1-score which are better than original metrics reported for the hybrid architecture. The results of class-wise prediction confirmed the effectiveness of the optimized model since F1-scores were equal to 0.9522, 0.8768, and 0.7619 for positive, negative, and neutral sentiments correspondingly. Even if the neutral sentiment was still the most difficult category to classify, the improved F1-score indicated that the optimization improved the performance and helped to mitigate the effects of the class imbalance in minority sentiments. As for the confusion matrices, the main diagonal showed that the class separation remained stable while the other errors concerned sentiment classes that were contextually ambiguous.

The results of ablation study have confirmed that using the full hybrid configuration contributed to achieving high F1-score. While BERT used alone produced an accurate result

of 0.9167, adding one more layer of either CNN (0.9196) or BiLSTM (0.9182) improved the performance. Finally, the full BERT-CNN-BiLSTM configuration obtained the best result of 0.9286 suggesting that the CNN and BiLSTM layers complemented each other providing local sentiment cues and bidirectional dependencies. The contextual comparison with reported FABSA baselines suggests that the proposed model is competitive, although the comparison cannot be treated as a strict leaderboard result because the baseline models were not reproduced under identical experimental conditions. However, in cross-dataset validation on the SemEval-2014 Restaurant data set, the model showed 0.7160 validation accuracy and 0.6547 weighted F1-score indicating weaker cross-dataset generalization ability. Overall, the results indicate that the proposed method was applicable for conducting context-aware sentiment analysis of e-commerce review data sets.

IV. Discussion

The results of the experiment show that the proposed BERT-CNN-BiLSTM architecture can effectively be used in sentiment classification of FABSA review cases. Specifically, the optimized version of the architecture proved to possess high validation scores of 0.9220 in terms of accuracy, 0.9225 in terms of weighted F1-score and 0.8636 in terms of macro F1-score outperforming the initial hybrid setup in all evaluated metrics. This is confirmed by the class-wise results of the validation, according to which the positive class received the highest F1 score while being followed by the negative and neutral classes. While neutral sentiment appears to be the most difficult to classify among other sentiments, its significant increase in F1-score from 0.591 to 0.7619 shows that optimized training has reduced the impact of class imbalance problem and increased the model's ability to recognize minority sentiment cases.

The ablation experiment supports the efficiency of the full-hybrid approach. In particular, BERT model provided high F1-score already, but the incorporation of either CNN or BiLSTM yielded additional improvements. Finally, BERT-CNN-BiLSTM produced the highest F1-score of 0.9286 suggesting that the addition of both CNN and BiLSTM branches provides additional advantage. The layers of CNN are helpful for recognizing phrase-level sentiment cues, while BiLSTM layers can model dependencies between the parts of the review. Therefore, the proposed architecture can be regarded as efficient both by general and by component-based assessment proving that contextual, local, and sequential models jointly enable more precise sentiment analysis.

Finally, the comparison with existing FABSA baseline models and additional testing on SemEval datasets demonstrate the pros and cons of the proposed model. It demonstrates competitive performance in FABSA review cases comparing with other baseline models of the framework; however, this comparison has to be made under contextual conditions since the leaderboard claim cannot be made due to potential differences in evaluations. However, zero-shot SemEval-2014 Restaurant testing revealed that it has low performance when compared with in-domain tests, which may be caused by the nature of domains. Still, the model retains the ability to correctly recognize polarity for positive and negative sentiments.

V. Limitations

There are several limitations in this study that need to be highlighted. First, although the preprocessing pipeline expands the review data using aspect–sentiment pairs, the aspect term itself is not explicitly appended to the review text during tokenization. Therefore, the current system should be understood as aspect-expanded sentiment classification rather than a fully aspect-conditioned sentiment model. Second, class imbalance remains an important challenge, particularly for the neutral class. Although stratified oversampling and class-weighted loss improve minority-class recognition, the neutral class still shows weaker performance than the positive and negative classes, especially under cross-dataset evaluation. Third, the model is evaluated primarily on the FABSA dataset, with additional zero-shot testing on SemEval-2014 Restaurant data. Broader validation across multiple product categories, domains, languages, and review platforms would be required to confirm the robustness and generalizability of the proposed approach. Finally, the study does not evaluate computational efficiency, inference time, model size, or deployment feasibility, which should be considered before applying the model in real-world e-commerce systems.

VI. Conclusion

This study proposed a hybrid BERT–CNN–BiLSTM model for context-aware sentiment classification of aspect-expanded e-commerce review data. The model combines BERT-based contextual representation learning, CNN-based local feature extraction, and BiLSTM-based bidirectional sequence modeling to capture both semantic and sequential patterns in review text.

Experimental results on the FABSA dataset show that the optimized hybrid model achieves 92.20% validation accuracy, 92.25% weighted F1-score, and 86.36% macro F1-score. The class-wise results further indicate improvement across all sentiment categories, with the strongest gain observed for the neutral class. This suggests that the optimized configuration improves class-balanced sentiment recognition rather than only improving performance on the dominant class.

The architecture-level ablation study confirms that the full BERT–CNN–BiLSTM model performs better than reduced variants using only BERT, BERT–CNN, or BERT–BiLSTM. In addition, the zero-shot evaluation on the SemEval-2014 Restaurant dataset shows that the model retains some ability to recognize sentiment polarity under domain shift, although neutral sentiment remains difficult to transfer across domains. Overall, the findings demonstrate that combining transformer-based contextual encoding with convolutional and recurrent components is an effective approach for sentiment classification in aspect-expanded e-commerce review data.

VII. Future Scope

Future research may extend the proposed sentiment classification framework by incorporating review helpfulness prediction as a separate downstream task or as part of a multi-task learning architecture. Such an extension would require datasets containing helpfulness labels or helpfulness votes, together with appropriate classification or regression-based evaluation metrics [21], [7], [24]. This would make it possible

to examine whether context-aware sentiment representations can also support the estimation of review usefulness.

Further improvement may also be achieved by making the model more explicitly aspect-aware. Although the present study expands review instances using aspect–sentiment pairs, the aspect term is not directly included in the tokenized input sequence. Future work may address this by structuring the input as a review–aspect pair, enabling the model to learn sentiment specifically toward the target aspect. This direction may be especially useful for reviews that contain mixed opinions about different product features [19], [6].

Future studies should also consider broader evaluation settings, including full test-set analysis, cross-validation, cross-domain evaluation, and comparisons across multiple benchmark datasets. Since the zero-shot evaluation on SemEval-2014 Restaurant data shows weaker transferability for the neutral class, additional experiments are needed to better understand the effects of domain shift, annotation differences, and neutral sentiment ambiguity [20]. Class imbalance should also be addressed more thoroughly using methods such as class-balanced loss, focal loss, aspect-level sampling, or targeted data augmentation for underrepresented sentiment classes [13], [20]. These strategies may further improve minority-class recognition, particularly for neutral sentiment. From a practical perspective, future work should also investigate inference speed, model size reduction, calibration, interpretability, and explainability so that the proposed approach can be adapted for real-world e-commerce review analysis.

VIII. Acknowledgment

We would like to express our sincere gratitude to our supervisor, Sifat Rahman Ahona, for the invaluable guidance, continuous support, and insightful feedback provided throughout the course of this research. We also extend our deepest appreciation to the Department of Computer Science and Engineering at the American International University-Bangladesh (AIUB) for providing the necessary facilities and fostering a conducive environment for our work.

References

[1] Y. Liu, J. Lu, J. Yang, and F. Mao, "Sentiment analysis for e-commerce product reviews by deep learning model of BERT-BiGRU-Softmax," *Mathematical Biosciences and Engineering*, vol. 17, no. 6, pp. 7819–7837, 2020.

[2] O. Bellar, A. Baina, and M. Ballafkih, "Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers," *Mathematics*, vol. 12, no. 15, p. 2403, 2024.

[3] A. A. Abohany, A. Shora *et al.*, "A hybrid deep learning approach for enhanced sentiment classification and consistency analysis in customer reviews," *Mathematics*, vol. 12, no. 23, p. 3856, 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/23/3856

[4] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications*, vol. 41, p. 3041–3046, 05 2014.

[5] J. Du, J. Rong, S. Michalska, H. Wang, and Y. Zhang, "Feature selection for helpfulness prediction of online product reviews: An empirical study," *PLOS ONE*, vol. 14, no. 12, p. e0226902, 2019.

[6] Y. Zhang, X. Li, and Y. Zhou, "A survey on online reviews in e-commerce: Classification, summarization, and sentiment analysis," *IEEE Access*, vol. 11, pp. 110 829–110 845, 2023.

[7] X. Li, Q. Li, and J. Kim, "A review helpfulness modeling mechanism for online e-commerce: Multi-channel cnn end-to-end approach," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023.

[8] Q. Tul, M. Ali, A. Riaz, A. Noureen, M. Kamranz, B. Hayat, and A. Rehman, "Sentiment analysis using deep learning techniques: A review," *International Journal of Advanced Computer Science and Applications*, vol. 8, 01 2017.

[9] A. M. van der Veen and E. Bleich, "The advantages of lexicon-based sentiment analysis in an age of machine learning," *PLoS ONE*, vol. 20, no. 1, p. e0313092, 2025.

[10] A. A. Maruf, F. Khanam, M. M. Haque, Z. M. Jiyad, M. F. Mridha, and Z. Aung, "Challenges and opportunities of text-based emotion detection: A survey," *IEEE Access*, vol. 12, 2024.

[11] M. M. Taye, R. Abulail, B. Al-Ifan, and F. Alsuhimat, "Enhanced sentiment classification through ontology-based sentiment analysis with bert," *Journal of Internet Services and Information Security (JISIS)*, vol. 15, no. 1, pp. 236–256, 2025.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[14] J. Shen, "Sentiment analysis of social media texts for aquatic product listed companies based on the BERT-CNN-BiLSTM-Att hybrid model," *Advances in Education, Humanities and Social Science Research*, vol. 13, pp. 182–185, 2025.

[15] L. Khan, A. Qazi, H.-T. Chang, M. Alhajlah, and A. Mahmood, "Empowering Urdu sentiment analysis: An attention-based stacked CNN-BiLSTM deep neural network with multilingual BERT," *Complex & Intelligent Systems*, 2024.

[16] A. Ba Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with turkish universities using twitter data," *Decision Analytics Journal*, vol. 11, p. 100473, 04 2024.

[17] R. Ahamad and D. Mishra, "Exploring sentiment analysis in handwritten and e-text documents using advanced machine learning techniques: a novel approach," *Journal of Big Data*, vol. 12, 01 2025.

[18] A. Angdresey, L. Sitanayah, and I. L. H. Tangka, "Sentiment analysis for political debates on youtube comments using bert labeling, random oversampling, and multinomial naive bayes," *Journal of Computing Theories and Applications*, vol. 2, no. 3, pp. 343–354, 2025.

[19] G. Kontonatsios, J. Clive, G. Harrison, T. Metcalfe, P. Sliwiak, H. Tahir, and A. Ghose, "Fabsa: An aspect-based sentiment analysis dataset of user reviews," *Neurocomputing*, vol. 562, p. 126867, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223009906

[20] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[21] H. Zureigat, O. Al-Qudah, and B. Alhijawi, "Predicting the helpfulness of online customer reviews," in *2023 International Conference on Information Technology (ICIT)*, Aug. 2023, pp. 686–689.

[22] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 08 2014.

[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 01 2018, pp. 328–339.

[24] S. Kim, S. Park, X. Li, and J. Kim, "A deep learning-based review helpfulness prediction system for online beauty products," *IEEE Access*, vol. PP, pp. 1–1, 01 2025.