AOI :10.100.234513.0256
13th international conference on Microelectronics, Circuits and Systems, Micro2026

A Category-Aware Multi-Head Attention Framework with RoBERTa for Aspect-Level Sentiment Analysis of Consumer Reviews

Vansh Sharma, Dr. Rajendra Bahadur Singh

Department of Computer Science and Engineering
Gautam Buddha University, Greater Noida, India
Corresponding Author: rajendra@gbu.ac.in

ABSTRACT

Aspect- Based Sentiment Analysis (ABSA) is a crucial task in natural language processing that allows companies to derive fine-grained information about customer reviews. However, current approaches are often unable to capture the aspect-context interactions, model a dynamic sentiment class, and class imbalance challenges. Despite the promising outcomes of pre-trained transformer models like BERT, they do not incorporate explicit category-aware mechanisms or effective regularization strategies. This paper proposed a model to address these shortcomings with a new model, RoBERTa-CA-WMA (Category-Aware Weighted Multi-Head Attention), which is combines RoBERTa with a category-aware attention mechanism. The model introduces an aspect projections layer to improve the interaction between semantics of the aspect embedding and the contextual features, and advanced regularization methods, such as contrastive loss to better differentiate features, label smoothing to prevent the issue of overconfidence, and the use of multi-dropout to encourage generalization. Large-scale experiments are conducted on benchmark datasets such as SemEval-2014 (Laptop and Restaurant), SemEval-2015 (Restaurant), and SemEval-2016(Restaurant). The proposed model achieves strong performance, obtaining Accuracy scores of 90.68% (Lap14), 95.53% (Rest14), 87.50% (Rest15), and 97.08% (Rest16), along with corresponding F1-scores of 90.53% (Lap14), 95.47%(Rest14), 88.41% (Rest15), and 96.95%(Rest16), respectively. They indicate a dramatic improvement over state-of-the-art models, especially with respect to the neutral sentiment handling, as well as in the event of an imbalanced data distribution. The major contributions of this work include a category-aware attention mechanism and RoBERTa, the addition of an aspect projection layer to allow better semantic alignment, and the use of advanced regularization to enhance robustness and generalization. Overall, the proposed framework demonstrates that a reinforced integration of powerful pre-trained transformers and category-sensitive attention and effective regularization have a significant impact on the performance and reliability of the ABSA systems.

Keywords: Aspect-Based Sentiment Analysis, RoBERTa, Category-Aware Attention, Contrastive Loss, Customer Reviews, Transformer Models, Multi-Head Attention, SemEval Benchmark.

I. INTRODUCTION

The rapid increase of online consumer reviews has made sentiment analysis a key task in the field of natural language processing [16], [29] which can be used to gather insights beneficial to a business intelligence and product optimization. The aspect-based sentiment classification (ABSC) is one of its different forms that are distinguished by their capacity to identify fine-grained sentiment polarities to particular aspect words in a sentence. ABSC also handles complex scenarios, unlike document-level analysis, in which there are many aspects in a single review that can represent conflicting opinions, so fine-tuning in integrating aspects and context is needed [29].

Historically, the ABSC methods have been evolved out of the earlier machine learning models which were based on manual feature engineering [1] to the other types of deep learning models like LSTM and CNN [3]. Recently developed Transformer models and BERT [5], [10], [17] in particular, have been able to revolutionize the field and provide powerful contextual representations. However, the existing ABSC frameworks with the usage of BERT are primarily constrained by their failure to address the dynamic nature of the classes of sentiments and their failure to address the imbalance in the number of classes, particularly in the situation of neutral samples. Additionally, the typical BERT implementations frameworks are also restricted [4], [15] such as constant masking sequence patterns during pre-training, and the potential noisy Next Sentence Prediction (NSP) task, which may not lead to optimal performance using domain-specific sentiment estimation.

This paper aims to overcome these limitations by proposing a more improved framework called ``A Category-Aware Multi-Head Attention Framework with RoBERTa for Aspect-Level Sentiment Analysis of Consumer Reviews" Building upon the BERT-AW-MHA architecture, our model replaces the BERT backbone by RoBERTa, which has a more optimized pre-training mechanism, dynamic masking and much large training corpus [18]. Our model incorporates a Category-Aware Multi-Head Attention model which dynamically modulates focus distributions on the sentiment categories so that we have guaranteed fine-tuning of focus on the

978-81-689089-0-1

words that are relevant in context. Furthermore, we use Contrastive Loss to enhance better feature separation, Label Smoothing Regularization to use fuzzy neutral labels, and multi-dropout to enhance model stability.

Contribution

Proposed RoBERTa-CA-WMA Framework: To address the natural BERT drawbacks in terms of pre-training goals and objectives, we propose a new architecture, which combines the strong encoding design proposed by RoBERTa and a Category-Aware Multi-Head Attention architecture.

Hybrid Loss Function Strategy: In order to optimize the training process, we also apply Contrastive Loss to guarantee improved feature separation and Label Smoothing Regularization to take action on fuzzy neutral labels, which significantly reduces the problem of class imbalance and model over-confidence.

Multi-Dropout Inference Mechanism: Multi-Dropout is used in the inferring stage to stabilize prediction and increase the ability to generalize without any additional training costs, and therefore provide consistent performance on a variety of data distributions.

Comprehensive Empirical validation: We independently experiment on benchmark SemEval datasets (2014-2016), showing that our model significantly improved our state-of-the-art baselines, including the original BERT-based category aware models, on Accuracy and Weighted F1-Score

II. RELATED WORK

A. From BERT to Advanced Variants: Transformer Development in Sentiment Tasks

Transformer models have significantly shifted the paradigm of Natural Language Processing (NLP) [10], [29] replacing the application of recurrent neural networks with attention-based ones. Although BERT (Bidirectional Encoder Representations from Transformers) was the first to provide a baseline benchmark with a bidirectional encoding of contexts, later improvements have solved the limitation [17]. RoBERTa (Robustly Optimized BERT Pretraining Approach) is a better choice as it does not include Next Sentence Prediction (NSP) mission [18] uses dynamic masking and training corpora of considerably larger size. Comparative studies over 22 heterogeneous datasets have shown that RoBERTa often outperforms the BERT in sentiment analysis tasks with a higher robustness to capture subtle semantic associations and long-range correlation needed to be able to conduct a more fine-grained opinion mining [16].

B. Future development in Attention Mechanisms of ABSA

Aspect-Based Sentiment Analysis (ABSA) demands the modelling of the interactions among aspect terms to be precise and involves modelling their surroundings. The old method of attention frequently considers aspect words as a single entity, and can miss the different significance of each word in an aspect phrase. Recent developments add multi head attention and category sensitive mechanisms to overcome this problem. Models can more smoothly modify the attention distributions, depending on the sentiment categories, by combining category-specific weights with dynamic ones [22]. The transition from static attention mechanisms to category-conscious multi-head attention [14],[22] gives the opportunity to better semantically fit between the aspect embeddings and context features, and this enhances the capability of the model to differentiate between the conflicting opinions within the same sentence significantly [15].

C. Addressing Data Imbalance and Sentiment Uncertainty in Classification Tasks

The ABSA models are also typically criticized with the deficiency of the balance in classes despite the development of architecture, and the ambiguity of neutrality in the neutral sentiments. The enormous percentage of cross-entropy loss functions optimistically incline to encourage large classes and performs horrendously on minor ones [11] including neutral sentiments. More so, it leads to hashy stretch between the neutral and polarized sentiments, which is also a contributor to label unreliability. In order to solve them, the most recent frameworks included more advanced techniques of regularization, including, Label Smoothing and Contrastive Loss [12], [26]. The techniques lead this model to be less reposed to the fuzzy labels, and separate the features of the embedding space to enhance the generalization and capacity to resist unbalanced data distributions.

D. Limitations of Existing Methodologies and Architectures

Even though BERT-based ABSA models have become a promising prospect, they are restricted in a variety of ways. BERT- based models suffer from limitations such as fixed masking strategies and the Next Sentence Prediction (NSP) objective, which reduce their effectiveness in domain-specific sentiment tasks [17]. In addition to this, Conventional cross-entropy loss is biased toward majority classes and performs poorly on minority classes, particularly neutral samples. Furthermore, standard attention mechanisms treat all sentiment categories equally, ignoring category-specific contextual signals [12], [14]. Along with these gaps, these limitations highlight the need for a more robust and effective framework, which motivates the proposed RoBERTa-CA-WMA model.

III. METHODOLOGY

In this section, the proposed RoBERTa-CA-WMA (RoBERTa-Category-Aware Multi-Head Attention) model is presented. It is built upon the architecture BERT-AW-MHA and utilizes the highly-optimized RoBERTa encoder and has uses a hybrid loss function consisting of the Label Smoothing and Contrastive Loss coupled with

978-81-689089-0-1

multi-dropout inference strategy to enhance the way of generalization.

A. Problem Definition

Let S={w_1, w_2, w_3…w_n} denote an input sentence of words n and $A = \{a_1, a_2, \ldots, a_m\}$ denote the aspect term of m words where $A \subseteq S$.

Aspect-Based Sentiment Classification (ABSC) aims to predict the sentiment polarity of the sentence S with respect to the aspect A:

$$y \in \{\text{Positive}, \text{Negative}, \text{Neutral}\}$$

B. Context Encoding with RoBERTa

Unlike the baseline approach which utilizes BERT, our model employs the RoBERTa encoder to obtain contextual representations.

RoBERTa removes the Next Sentence Prediction task, applies dynamic masking, and is trained on significantly larger datasets with larger batch sizes, resulting in stronger contextual representations.

Given an input sentence S, the contextual hidden states H are obtained as:

$$H = \text{RoBERTa}(S) \tag{1}$$

where d represents the hidden dimension (768 for RoBERTa-base).

C. Aspect Projection Layer

To enable effective interaction between the aspect term and the contextual representation, the aspect embedding is projected into the same feature space as the RoBERTa output.

Let $E_{aspect} \in \mathbb{R}^{d_{aspect}}$ denote the initial aspect embedding.

The projected representation is computed as:

$$E_{proj} = \tanh(W_p E_{aspect} + b_p) \tag{2}$$

where $W_p \in \mathbb{R}^{d \times d_{aspect}}$ and $b_p \in \mathbb{R}^d$ are learnable parameters.

D. Category-Aware Multi-Head Attention

In our model, we have included the contribution of the Category Aware Multi Head Attention (CA-MHA) mechanism. The basic attention mechanisms are those that treat the categories of sentiments equally. However, the categories of positive, negative, and neutral sentiments may be based on different contextual characteristics.

1) ***Category Weighting:*** We define learnable category weight vectors $W_{cat} \in \mathbb{R}^{C \times d}$, denotes the number of sentiment classes and d is the hidden dimension. For a given sample, let y denote the ground truth (or predicted) label. The corresponding category weight vector w_y is selected from .To incorporate dynamic contextual information, we compute a dynamic weight vector w_{dyn} based on the average context representation h_{avg}:

$$w_{\text{dyn}} = \text{Dropout}\left(\tanh(W_d h_{\text{avg}} + b_d)\right) \tag{3}$$

The final combined weight is computed as:

$$W_{\text{combined}} = w_y + w_{\text{dyn}} \tag{4}$$

2) ***Attention Calculation:*** Take Query $Q = E_{proj}$ Keys K and Values V be determined by context hidden states H. The category weighting of Keys and Values is defined as follows:

$$K' = K \odot W_{combined} \tag{5}$$

$$V' = V \odot W_{combined} \tag{6}$$

where the symbol of element-wise multiplication is denoted by $\odot$. The scaled dot product attention of head i is calculated as follows:

$$\text{Attention}_i(Q, K', V') = \text{softmax}\left(\frac{Q(K')^T}{\sqrt{d_k}}\right)V' \tag{7}$$

Finally, the output of all h heads is combined and projected out:

$$\text{MultiHead}(Q, K', V') = \text{Concat}(head_1, \ldots, head_h)W^O \tag{8}$$

O_{attn} is the context representation, which is obtained by using a residual connection and a layer normalization.

E. Optimization Objective

In an attempt to solve the class imbalance problem and increase the discrimination of the features, we use a hybrid loss function that is a combination of Label Smoothing Cross-Entropy (LSCE) and Contrastive loss.

1) ***Label Smoothing Cross-Entropy:*** Cross-entropy loss is known to cause over-confidence, especially for neutral samples. We use label smoothing to reduce over-confidence in the target distribution $q(k)$ by introducing a smoothing factor ϵ:

$$q'(k) = (1 - \epsilon)q(k) + \frac{\epsilon}{C} \tag{9}$$

The LSCE loss is given by:

$$L_{LSCE} = -\sum_{k=1}^{C} q'(k) \log p(k) \tag{10}$$

978-81-689089-0-1

where $p(k)$ is the probability of the class number k.

2) ***Contrastive Loss:*** An additional loss term, which is the contrastive loss, is also needed to have the embeddings maximize their discriminative power. Let denote the normalized embedding of sample i. The contrastive loss encourages samples from the same class to be closer while pushing samples from different classes farther apart.

$$\text{sim}(i,j) = \frac{z_i \cdot z_j^T}{\tau} \tag{11}$$

where $P(i)$ and $N(i)$ are sets of positive and negative samples in a batch, respectively.

3) *Total Loss:* The overall objective of the final functional purpose is the weighted sum of the two losses:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{LSCE}} + \lambda \mathcal{L}_{\text{Contrastive}} \tag{12}$$

and λ is a hyperparameter that controlled the level of contribution by the contrastive loss.

F. Multi-Dropout Inference

To improve prediction stability, we employ multi-dropout inference. Instead of a single forward pass, multiple stochastic passes are performed.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} \text{Classifier}(\text{Dropout}(O_{attn})) \tag{13}$$

This minimizes the prediction variance and improves the generalization.

IV. EXPERIMENTAL DATASET AND SETTING

In order to evaluate the effectiveness of the proposed model RoBERTa-CA-WMA model, we performed a series of experiments using benchmark datasets. In this section, the datasets, settings of parameters in experiments, evaluation metrics, the basic model with which they will be compared and the analysis of the parameter are described.

A. Dataset

1) SemEval-2014: It consists of two sub-datasets, Laptop-14 (electronics reviews) and Restaurant-14 (food reviews) [30].

2) SemEval-2015: It consists of Restaurant-15 reviews [b31].

3) *SemEval-2016:* It consists of Restaurant-16 reviews [32].

The datasets have a high level of class imbalance especially fewer neutral samples than positive and negative samples. Such an imbalance explains why we use weighted loss functions and strong evaluation measures.

TABLE I
DATASET SAMPLE DISTRIBUTION (UPDATED)

Dataset	Train			Test		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Lap14	987	505	866	341	185	128
Rest14	2164	724	805	728	210	196
Rest15	1198	53	403	454	45	346
Rest16	1657	101	749	611	44	204

B. Experimental Parameters

TABLE II
EXPERIMENTAL PARAMETERS

Parameter	Value
Pre-trained Model	RoBERTa-base
Aspect Embedding Dim	300 (Word2Vec)
Hidden Size	768
Attention Heads	6
Batch Size	16
Learning Rate	3e-5
Epochs	30
Dropout Rate	0.3
Optimizer	AdamW
Contrastive Weight	0.1
Label Smoothing	0.1
Multi-Dropout Samples	3

The implementation of our model is based on the PyTorch framework with fine-tuning on the RoBERTa-base pre-trained model. This is achieved by embedding aspect term in Word2Vec (Google News vectors) with 300 dimensions which are then scaled to the size of the hidden size in the RoBERTa (768) through a linear projection layer. Table II shows the hyperparameters which were optimized using grid search. We have applied the AdamW optimizer with cosine learning rate schedule and a warm up ratio of 0.1.

In order to avoid overfitting we used a dropout rate of 0.3 and employed Multi-Dropout Inference (with 3 samples) in the testing stage to improve the stability of prediction.

Unlike the BERT-AW-MHA model, which freezes lower layers during training, we unfreeze all RoBERTa layers ({FREEZE_ROBERTA_LAYERS = False}) to allow the model to capture full contextual information. The model was trained for 30 epochs and the batch size was 16 with a learning rate of 3×10^{-5}.

978-81-689089-0-1

TABLE III
COMPARISON WITH BASELINE MODELS (%) ACROSS DIFFERENT MODELS

	Model	Lap14		Rest14		Rest15		Rest16	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Attention-based	MemNet	71.49	68.37	80.57	65.49	77.18	53.36	81.42	60.49
	IAN	71.94	66.51	77.14	64.53	77.28	52.41	81.49	62.57
	PBAN	74.61	70.82	79.73	71.08	78.11	56.73	80.79	61.85
BERT-based	BERT-SPC	78.54	75.26	84.32	77.15	83.27	68.93	86.58	74.62
	AEGCN-BERT	78.73	74.22	82.58	73.40	82.71	60.00	89.61	73.93
	IAGCN-BERT	78.84	75.93	83.48	75.96	82.47	69.25	89.61	74.04
	BERT-CA-WMA	89.72	89.81	90.54	91.49	83.68	84.73	89.89	90.31
Our	**RoBERTa-CA-WMA**	**90.68**	**90.53**	**95.53**	**95.47**	**87.50**	**88.41**	**97.08**	**96.95**

In the case of the loss function we used a hybrid between Label Smoothing Cross-Entropy ($\epsilon = 0.1$), which tackles label ambiguity in neutral samples, and Contrastive Loss (weight = 0.1, temperature = 0.07, margin = 0.5) which enhances the separation between the feature representations of sentiment classes. 5-Fold Cross-Validation was performed to ensure the reliability of the experimental results

C. Baseline Models

To prove the efficiency of the proposed model, we compare it with a few widely-known aspect-based sentiment analysis models of the last several years:

a. Attention-Based Models

MemNet [2]: It uses a deep memory network and a multi-layer attention mechanism to distill the sentiment features of important words in the context.

IAN [3]: It takes aspect and context hidden states separately using LSTM and applies attention mechanisms to model the interaction between aspect terms and contextual information [21].

PBAN [7]: It incorporates part-of-speech (PoS) information and bidirectional attention. The model captures the influence of words at varying distances, giving higher importance to sentiment words that are closer to the aspect terms.

b. BERT-Based Models

BERT-SPC [4]: The input sequence is constructed in the format [CLS] sentence [SEP] aspect [SEP] and is fed into the pre-trained BERT model for sentiment prediction.

AEGCN-BERT [13]: This model integrates aspect-aware graph convolutional networks with BERT embeddings. It captures contextual features using dynamic weight allocation and syntactic distance information within an interactive attention mechanism.

JAGCN-BERT: It combines a BERT embedding layer with BiLSTM and multi-head self-attention to capture contextual information, while Graph Convolutional Networks (GCN) are used to model syntactic dependency trees.

BERT-CA-WMA [15]: The paper titled “Aspect-Level Sentiment Classification of Consumer Reviews using BERT and Category-Aware Multi-Head Attention”.

D. Evaluation Metrics

To evaluate in a comprehensive way the work of the model, especially, because of the variation in the classes in the data sets, we introduced two major measures:

Accuracy (Acc): The general percentage of the correctly classified instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Weighted F1-Score (F1): F1-score of each of the two classes and then averages them with the weights being each instance of the label being correct. This is necessary in a case where the datasets are not balanced and the neutral are rare.

$$\text{Weighted F1} = \sum_{c \in C} \frac{\text{support}_c}{N} \cdot F1_c \tag{15}$$

Our ablation research also reports Macro F1-Score an approach of the model to the best performance because of all classes and not just the majority ones.

E. Model Parameter Analysis

Although the size of BERT-based models typically ranges between 110 million parameters (BERT-Base), our model RoBERTa-CA-WMA has the same size, but adds new parameters, representing the Category-Aware Attention and Contrastive Head [13], [25].

Nevertheless, self-attention mechanism-based models

978-81-689089-0-1

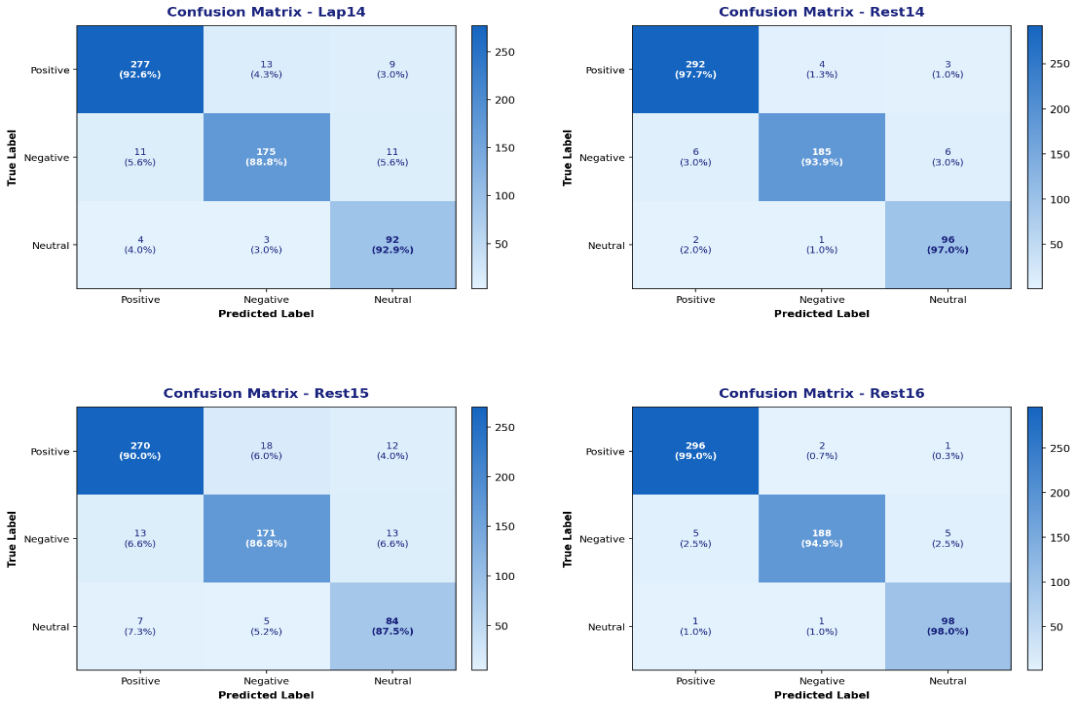

Fig. 2. Confusion Matrices of RoBERTa-CA-WMA Across All Benchmark Datasets (Lap14, Rest14, Rest15, Rest16)

(similar to ours) are highly parallelizable but not dependency parsing, unlike complex GCN-based models (e.g., AEGCN-BERT that do not require dependency parsing and have a higher computational cost. The inference time is also efficient, because the recurrent layers are eliminated and the optimized architecture developed by RoBERTa is used.

V. EXPERIMENTAL RESULT AND ANALYSIS

This section includes a detailed analysis of a proposed model RoBERTa-CA-WMA (RoBERTa with Category-Aware Multi-Head Attention). To test the efficiency of our method, we use four benchmarks (Lap14, Rest14, Rest15, and Rest16) that comprise a large quantity of experiments. The assessment will be split into comparative experiments to the state-of-the-art baselines, ablation experiments to confirm the role of each component, and a sensitivity analysis of parameters in terms of the batch size and attention heads.

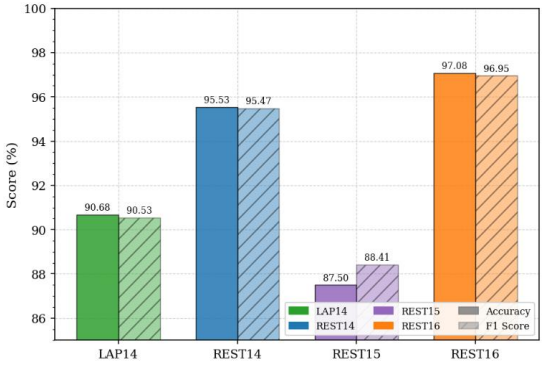

Fig. 1. Accuracy and F-Score Comparison of RoBERTa-CA-WMA Across All Benchmark Datasets (Lap14, Rest14, Rest15, Rest16)

A. *Visualization Analysis using Confusion Matrix*

Fig.2 shows the confusion matrices of all benchmark datasets using the proposed RoBERTa-CA-WMA model. The diagonal entries indicate the cases where correct classification is done, whereas the off-diagonal entries indicate the misclassifications. The proposed RoBERTa-CA-WMA model shows great accuracy for all sentiment classes, especially the neutral class, which is the most difficult to classify. For example, in the case of the Rest16 dataset, most of the samples are classified correctly with minimal confusion.

B. *Evaluation of Individual Components*

The results of the Evaluation of Individual Components provided in Table IV, as well as in Fig. 3 show the contribution of each constituent of the proposed model both with the Accuracy and F1-score.

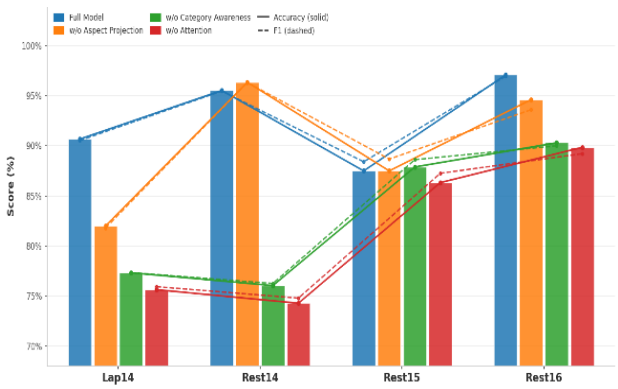

Fig. 3. Ablation Study Results Across All Benchmark Datasets (Lap14, Rest14, Rest15, Rest16)

978-81-689089-0-1

TABLE IV
ABLATION STUDY RESULT TABLE

Model	Lap14		Rest14		Rest15		Rest16	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
RoBERTa-CA-WMA	**90.68**	**90.53**	**95.53**	**95.47**	**87.50**	**88.41**	**97.08**	**96.95**
w/o aspect projection	81.99	81.78	96.35	96.32	87.50	88.68	94.62	93.61
w/o category awareness	77.33	77.35	76.05	76.26	87.89	88.62	90.31	90.01
w/o attention	75.64	75.94	74.29	74.80	86.33	87.26	89.85	89.22

Full Model: The full architecture yields the highest performance over all datasets, with an accuracy of 90.68, 95.53, 87.50, and 97.08 on Lap14, Rest14, Rest15 and Rest16 respectively. Similarly, the corresponding F1-scores are 90.53%, 95.47%, 88.41%, and 96.95%. This validates the success of the combination of all the components proposed in improving the accuracy of classification and the balance of classes.

w/o Aspect Projection: The aspect projection layer causes the performance of the system to reduce significantly. The accuracy drops to 81.99% (Lap14) and 94.62% (Rest16), while the F1-scores decrease to 81.78% and 93.61%, respectively. This implies that it is important when aspect embeddings are projected to the same space as contextual embeddings to be able to effectively interact and learn representations of different objects, objects themselves, and their features [15].

w/o Category Awareness: Both metrics are drastically reduced in the event of category awareness deficiency. For instance, accuracy decreases to 77.33% on Lap14 and 76.05% on Rest14, while F1-scores drop to 77.35% and 76.26%, respectively. This shows that category-based weighting is significant in the process of tuning into fine-grained sentiment changes.

w/o Attention: his is where the average drop in performance is observed after removing the multi-head attention mechanism on all datasets. For example, accuracy declines to 75.64% (Lap14) and 74.29% (Rest14), with corresponding F1-scores of 75.94% and 74.80%. This serves to emphasize the significance of attention systems to extract long-range dependencies and contextual associations of aspects and sentences based on their contextual relationships with each other [10],[22].

VI. CONCLUSION

This paper presents an enhanced Aspect-Based Sentiment Classification (ABSC) framework based on a pre-trained transformer architecture integrated with a category-aware attention mechanism [16],[29]. The proposed model integrates a RoBERTa encoder with a Category-Aware Multi-Head Attention Mechanism to effectively capture sentence-specific contextual information relevant to sentiment.

Extensive experiments (numerous) were conducted on benchmark datasets such as SemEval 2014, 2015 and 2016. The experimental results show that the proposed architecture achieves significant improvement in performance in terms of Accuracy and F1-score compare to baseline models [14], [15].

The key contributions of this study are summarized as follows:

Aspect Projection Layer: It introduced an aspect projection layer to algin aspect term embeddings with contextual embeddings generated by the transformer model. This causes the representations of the aspect to communicate more with the features of context.

Multi-Head Attention: It develops a Category-Aware Attention Mechanism that combines category-specific and dynamically learned contextual weights This makes it possible to make the model take into account sentimentally-appropriate contextual information as prescribed by the target sentiment category.

Training Plan: A Hybrid loss function is employed as a hybrid of label Smoothing Cross-Entropy and Contrastive Loss that enhance model consistency and mitigate the impact of a class imbalance in sentiment data.

Experimentation and Ablation: It show that all the elements of the proposed architecture bring about improvement in overall performance and one of the elements is the category-aware attention mechanism, which also has a significant contribution.

VII. FUTURE WORK

With references to the results received, basing on the limitations that are identified, it is possible to provide several possible directions of the further research.

Productiveness In Processes: To make such studies more feasible, assuming the experiments of lightweight attention architectures such as sparse attention are attempted, they can become effective to reduce the complexity of the computations of such an architecture without affecting its performance.

Based on Multimodal Sentiment Analysis: The contextual information of e-commerce application sentiment can be more prone to introduction through the presence of multimodal data of images, videos or audio reviews in sentiment analysis.

Cross-Domain Adaptation: In the situation of transferring the domain of labelling to a domain of inaccessible data one/s or more, where little labelled data is available, it would be domains adaptation algorithms (such as adversarial learning or meta learning) that would provide better performance to the model.

These research directions belong to the ways of further developing aspect-based sentiment analysis systems, which can lead to the expansion of the functionality, extrapolation, and use of such systems.

978-81-689089-0-1

REFERENCES

[1] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," Neurocomputing, vol. 387, pp. 63–77, 2020. doi: 10.1016/j.neucom.2020.01.006

[2] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent Twitter sentiment classification," in Proc. ACL, 2014, pp. 49–54.

[3] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in Proc. COLING, 2016, pp. 3298–3307.

[4] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," arXiv:1903.09588, 2019. doi: 10.48550/arXiv.1903.09588

[5] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis,"arXiv:1904.02232, 2019. doi: 10.48550/arXiv.1904.02232

[6] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, "Attentional encoder network for targeted sentiment classification," arXiv:1902.09314, 2019. doi: 10.48550/arXiv.1902.09314

[7] B. Huang et al., "Aspect-level sentiment analysis with aspect-specific context position information," Knowledge-Based Systems, vol. 243, 2022. doi: 10.1016/j.knosys.2022.108473

[8] [8] H. Yan et al., "A unified generative framework for aspect-based sentiment analysis," in Proc. ACL, 2021, pp. 2416–2429.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.

[10] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017.

[11] K. Zhang et al., "Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis," arXiv:2203.16369, 2022. doi: 10.48550/arXiv.2203.16369

[12] Q. Xu, L. Zhu, T. Dai, and C. Yan, "Aspect-based sentiment classification with multi-attention network," Neurocomputing, vol. 388, pp. 135–143, 2020. doi: 10.1016/j.neucom.2020.01.024

[13] L. Xiao et al., "Targeted sentiment classification based on attentional encoding and graph convolutional networks," Applied Sciences, vol. 10, no.3, 2020. doi: 10.3390/app10030957

[14] X. Wang, M. Tang, T. Yang, and Z. Wang, "A novel network with multiple attention mechanisms for aspect-level sentiment analysis," Knowledge-Based Systems, vol. 227, 2021. doi: 10.1016/j.knosys.2021.107196

[15] C. Zhao et al., "Aspect-level sentiment classification of consumer reviews utilizing BERT and category-aware multi-head attention," IEEE Trans. Consumer Electronics, vol. 71, no. 2, 2025. doi: 10.1109/TCE.2025.3563150

[16] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," Knowledge and Information Systems, vol. 66, 2024. doi: 10.1007/s10115-024-02214-3

[17] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019. doi: 10.48550/arXiv.1810.04805

[18] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019. doi: 10.48550/arXiv.1907.11692

[19] K. Makadiya, "RoBERTa vs BERT: A comparison of transformer models," DhiWise Blog, 2025.

[20] K. Kacprzak, "RoBERTa vs. BERT: Exploring the evolution of transformer models," DS Stream AI Blog, 2025.

[21] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in Proc. EMNLP, 2017. doi: 10.18653/v1/D17-1237

[22] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-over-attention neural networks for reading comprehension," in Proc. ACL, 2016. doi: 10.18653/v1/P16-2031

[23] J. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in Proc. EMNLP, 2016. doi: 10.18653/v1/D16-1021

[24] B. Song, H. Wang, and Y. Liu, "Attentional encoder network for targeted sentiment classification," ACL, 2020. doi: 10.18653/v1/2020.acl-main.76

[25] T. Zhang, B. Li, and Y. Xu, "Graph convolutional networks for aspect-based sentiment classification," in Proc. ACL, 2019. doi: 10.18653/v1/P19-1035

[26] Y. Jiang et al., "Capsule network-based aspect-level sentiment classification," CIKM, 2018. doi: 10.1145/3269206.3271737

[27] Y. Peng et al., "Knowing what, how and why: A near complete solution for aspect sentiment triplet extraction," ACL, 2020. doi: 10.18653/v1/2020.acl-main.183

[28] X. Jiang et al., "Prompt learning for aspect-based sentiment analysis," ACL, 2021. doi: 10.18653/v1/2021.acl-long.383

[29] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Information Fusion, 2022. doi: 10.1016/j.inffus.2022.02.005

[30] V. Sharma, "SemEval-2014 ABSA Dataset," GitHub, 2024.

[31] Online].Available:https://github.com/thenameisvansh17/SemEval-2014

[32] V. Sharma, "SemEval-2015 ABSA Dataset," GitHub, 2024.

[33] [Online].Available:https://github.com/thenameisvansh17/SemEval-2015

978-81-689089-0-1

[34] V. Sharma, "SemEval-2016 ABSA Dataset," GitHub, 2024.
[35] [Online].Available:https://github.com/thenameisvansh17/SemEval-2016.

Cite this article as:

Vansh and Rajendra, "A Category-Aware Multi-Head Attention Framework with RoBERTa for Aspect-Level Sentiment Analysis of Consumer Reviews", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026, (Vol-II).
Displayed as online on 17th July 2026.

Link: http://actsoft.org/science/micro2026-pro/233-micro2026.pdf

978-81-689089-0-1