Enhanced Aspect-Based Sentiment Analysis with Integrated Category Extraction for Instruct-DeBERTa

Dineth Jayakody[1,*], **Koshila Isuranda**[1,*], **A V A Malkith**[1,*], **Nisansa de Silva**[2,§], **S R Ponnamperuma**[3,¶], **G G N Sandamali**[1,†], **K L K Sudheera**[1,†], **K G Sarathchandra**[3,¶]
[1]University of Ruhuna, Sri Lanka, [2]University of Moratuwa, Sri Lanka, [3]Emojot Inc.
[*]{jayakody_ds_e21,isuranda_mak_e21,malkith_ava_e21}@eng.ruh.ac.lk,
[§]NisansaDdS@cse.mrt.ac.lk, [†]{nadeesha,kushan}@eie.ruh.ac.lk, [¶]{sachintha,kashnika}@emojot.com

Abstract

Aspect-Based Sentiment Analysis (ABSA) has seen significant advancements with the introduction of Transformer-based models, which have reshaped the landscape of Natural Language Processing (NLP) tasks. This paper introduces enhancements to the `Instruct-DeBERTa` model which is one of the leading ABSA models for ABSA. It takes a hybrid approach combining the strengths of `InstructABSA` for Aspect Term Extraction (ATE) and `DeBERTa-V3-baseabsa-V1` for Aspect Sentiment classification (ASC). In this work, we enhance `Instruct-DeBERTa` by introducing category classification through a cosine similarity-based method, comparing aspect embeddings with predefined categories. Also for `InstructABSA` and `DeBERTa-V3-baseabsa-V1`, we investigate different configurations by adding a linear layer followed by ReLU activation, incorporation of regularization and optimization of attention heads. These modifications were tailored specifically for the data sets in the hospitality domain. Our empirical evaluations, run on diverse datasets, have shown that these enhancements significantly raise the performance of `Instruct-DeBERTa` for hospitality domain datasets.

1 Introduction

The growing interest in NLP makes ABSA an important building block for sentiment detection and investigation using textual information (Mudalige et al., 2020; Rajapaksha et al., 2020). Unlike traditional approaches to sentiment analysis, where just the estimate of polarity value was estimated, ABSA focuses on fine-grained opinions expressed on some features or attributes offered by products or services (Rajapaksha et al., 2021; Jayasinghe et al., 2021). This is especially important for any business wishing to understand customer feedback better and improve products and services based on the overall opinion of the consumers.

It was only in the most recent years that one witnessed substantial progress in machine and deep learning applied to ABSA methodologies (Rajapaksha et al., 2022; Samarawickrama et al., 2022). Early lexicon-based approaches failed to properly account for context and ambiguity, while later-introduced machine learning models were most of the time heavily reliant on manual feature engineering and lacked generalization across domains. Significant progress has been associated with its application, especially through models such as recurrent neural networks, long short-term memory networks, and convolutional neural networks. But still, capturing long-term dependencies and complex syntactic structures effectively remains hard.

Transformer-based architectures, most notably exemplified by `BERT`, revolutionized the field by using attention mechanisms to capture contextual relationships from all directions within a sentence. Having advanced their ability to further comprehend complex linguistic patterns and relations, these models set new records on many NLP tasks. In this line of research, state-of-the-art models that emerge are `InstructABSA` for ATE and `DeBERTa-V3-baseabsa-V1` for ASC. The work presented by Jayakody et al. (2024b) introduces `Instruct-DeBERTa` — a hybrid model that combines the best of `InstructABSA` (Scaria et al., 2024) in ATE with those of `DeBERTa-V3-baseabsa-V1` (Yang et al., 2023, 2021) in ASC. The model was constructed to perform the joint task of aspect extraction and sentiment polarity detection within a single pipeline. Evaluation across the SemEval 2014-2016 restaurant reviews (Res-14, Res-15, and Res16)and the SemEval 2014 laptop dataset (Lap-14), has demonstrated that `Instruct-DeBERTa` is better by quite a margin than any other model in accuracy and robustness and is hence likely state-of-the-art for the joint task of ATE and ASC.

However, there are always some aspects that

Model	F1 Score (%)					
	Res-14		Res-15		Res-16	
	ATE	ASC	ATE	ASC	ATE	ASC
InstructABSA (Scaria et al., 2024)	92.10	–	76.64	–	80.32	–
DeBERTa-V3-base-absa-v1.1 (Yang et al., 2023, 2021)*	–	90.94	–	89.55	–	83.71
DeBERTa-V3-base-absa-v1.1-Improved version	–	**91.62**	–	**86.79**	–	**85.88**
Instruct-DeBERTa (Single task)*	91.39	88.63	75.13	81.26	77.79	79.35
Instruct-DeBERTa-Improved version (Single task)	**91.39**	**89.22**	**75.13**	**81.14**	**77.79**	**80.61**
Instruct-DeBERTa (Joint task)*	80.78		–		–	
Instruct-DeBERTa-Improved version (Joint task)	**81.64**		**68.93**		**72.23**	

Table 1: F1 scores for the selected models individually and when pipe-lined. Note*: These F1 scores were taken from Jayakody et al., 2024b.

remain quite underdeveloped in the case of Instruct-DeBERTa. In this work, we make a few substantial improvements beyond the base model. We include a component for category classification with cosine similarity to classify the extracted aspects by comparing them with the pre-trained embeddings of categories. This is then plotted on a Voronoi diagram to clearly and intuitively provide insight into how the aspects are spread across different categories. Furthermore, we did extensive hyper-parameter tuning and architectural changes of our model with availabl for especially on the DeBERTa-V3-baseabsa-V1 model—to ensure that our trained model works most effectively on the hospitality domain. This also increases the capacity to classify sentiment polarities accurately. These numerous innovations further open up the horizons of ABSA in order to have a more detailed and precise model for the analysis of customer feedback.

2 Background

Recent studies have explored advanced methodologies to enhance the efficiency and scalability of ABSA models. These include using the Quantized Low-Ranking Adaptation (QLoRA) (Dettmers et al., 2023) approach to Llama 2 (Touvron et al., 2023) fine-tuning, utilizing the SETFIT (Tunstall et al., 2022) framework for few-shot learning, and implementing FAST_LSA_T_V2 (Yang and Li, 2024) within the PyABSA (Yang et al., 2023) framework. Among them, the best result was produced by the FAST_LSA_T_V2 model with 87.6% and 82.6% on the Res-14 and Lap-14 datasets, respectively. None of these models outperformed the reported LSA+DeBERTa-V3-Large (Yang and Li, 2024) model by the accuracy of 90.33% and 86.21% on the same datasets (Jayakody et al., 2024a). This study mainly focused on single-task ABSA in the effort of establishing a hybrid model for performance in certain domains such as restaurants and laptops.

In general, there are two main underlying ABSA subtasks: Aspect Term Extraction and Aspect Sentiment Classification. Transformer-based models have significantly advanced the performance of these tasks. Very recently, the authors of Jayakody et al., 2024b have therefore proposed an ABSA pipeline chain based on Transformer-based models that will automatically extract aspects and perform the sentiment analysis in the text data.

In the present review, the best model performance was identified for each of the subtasks. However, the instructABSA has performed the best on the ATE task so far, with 92.10% F1 on the Res-14 dataset, outperforming every other model that also had equally very good performance for all other datasets such as Res-15, Res-16, and Lap-14, showing strong generalization capability across domains. Among these, DeBERTa-V3-base-absa-v1 was the best in the general ASC task, showing the highest F1 score on all datasets. For example, the Res-14 dataset alone recorded 90.94%. Its performance was considered quite good for all datasets across Res-15, Res-16, and Lap-14, which were from different domains. Based on these results, a hybrid model, termed Instruct-DeBERTa, was proposed, consisting of a pipelined combination of the InstructABSA model for ATE and the DeBERTa-V3-base-absa-v1 model for ASC, where the benefits of both models are sought to be utilized in accomplishing the joint ABSA task.

Instruct-DeBERTa demonstrates strong performance across various sentiment classification tasks, with most of the extracted and classified aspects achieving high F1 scores, underscoring the model's precision and stability. As illustrated in Table 1, although there was a slight decrease in some F1 scores due to the pipelining process referenced in Jayakody et al., 2024b, the hybrid model's overall performance remained resilient. Particularly, the model performs exceptionally well in the joint task, achieving pair extraction F1 scores of 80.94%

for the Lap-14 dataset and 80.78% for the Res-14 dataset. These results underscore the model's durability and efficacy by showing that it can attain higher accuracy than what has been previously reported for these datasets.

3 Methodology

Under this section, we discuss on optimizing the performance of `Instruct-DeBERTa` for enhanced efficiency in ABSA in the hospitality domain. Rs-14, Res-15, and Res-16 are the main data sets that we utilize in the analysis to focus on this domain. More importantly, a new mechanism for category classification is introduced, and the model architecture parameters are fine-tuned. The overall structure of our model is shown in Fig. 1.

First, we developed a categorization classification method through which the identified aspects were allocated to the established categories, using a cosine similarity-based methodology. Further elaboration of this development will enhance the accuracy of analysis and allow better structuring and interpretation for the sentiments associated with these aspects. This is undertaken for visualization using Voronoi diagrams in order to exactly understand how such aspects distribute within the categories in a very clear and intuitive way. Based on this work, we fine-tuned some additional model architecture parameters for the dropout rates, the attention mechanism, layer normalization, and several others, within the `DeBERTa-V3` and `InstructABSA` models. This was done to further compress more improvements into the model with respect to accuracy and robustness in the classification of aspects and sentiment polarity.

3.1 Integrating aspect categorization

In order to improve the `Instruct-DeBERTa` model, we embedded aspect category separation within the domain of sentiment analysis. The model categorizes each aspect term identified within a sentence into predefined categories, using an embedding-based similarity approach. Additionally, we visualized the relationships between these aspects and their categories using t-SNE dimensionality reduction and Voronoi diagrams. This whole process was explicitly done without training the model on a certain dataset that would contain both aspects and categories, but categorization has been purely based on similarities between embeddings.

The core functionality of the model is to categorize aspect terms into specific categories. This was achieved using an embedding-based method where each aspect term is embedded into a high-dimensional vector space using `GIST-Embedding-v0` (Solatorio, 2024). This model was chosen since it was the best performing embedding model with the least amount of model parameters and embedding dimensions. This addition of the embedding model made the collective hybrid model `Instruct-DeBERTa` a single triple task model consisting of `InstructABSA`, `DeBERTa-V3` and `GIST-Embedding-v0`. The aspect term is then categorized based on its similarity to predefined category embeddings.

The categorization process is mathematically formalized as follows:

$$\mathbf{e}_{\text{aspect}} = \text{Encode}(\text{aspect}) \tag{1}$$

Where:

- $\mathbf{e}_{\text{aspect}}$ represents the embedding of the aspect term, obtained using the embedding model's encode function.

The similarity between the aspect embedding and each category embedding is calculated using the cosine similarity function:

$$\text{CS}(\mathbf{e}_{\text{aspect}}, \mathbf{e}_{\text{category}}) = \frac{\mathbf{e}_{\text{aspect}} \cdot \mathbf{e}_{\text{category}}}{\|\mathbf{e}_{\text{aspect}}\| \|\mathbf{e}_{\text{category}}\|} \tag{2}$$

Where:

- CS stands for Cosine Similarity
- $\mathbf{e}_{\text{category}}$ is the embedding of a predefined category.
- $\cdot$ denotes the dot product, and $\|\cdot\|$ represents the vector norm.

The aspect term is assigned to the category with the highest average cosine similarity score:

$$\text{Best Category} = \arg \max_{\text{category}} \frac{1}{n} \sum_{i=1}^{n} \text{CS}(\mathbf{e}_{\text{aspect}}, \mathbf{e}_{\text{category}}) \tag{3}$$

Where:

- n represents the number of embeddings per category.

This approach ensures that each aspect term is grouped with the category that it is most semantically aligned with, according to the vector representations learned by the embedding model. Also,

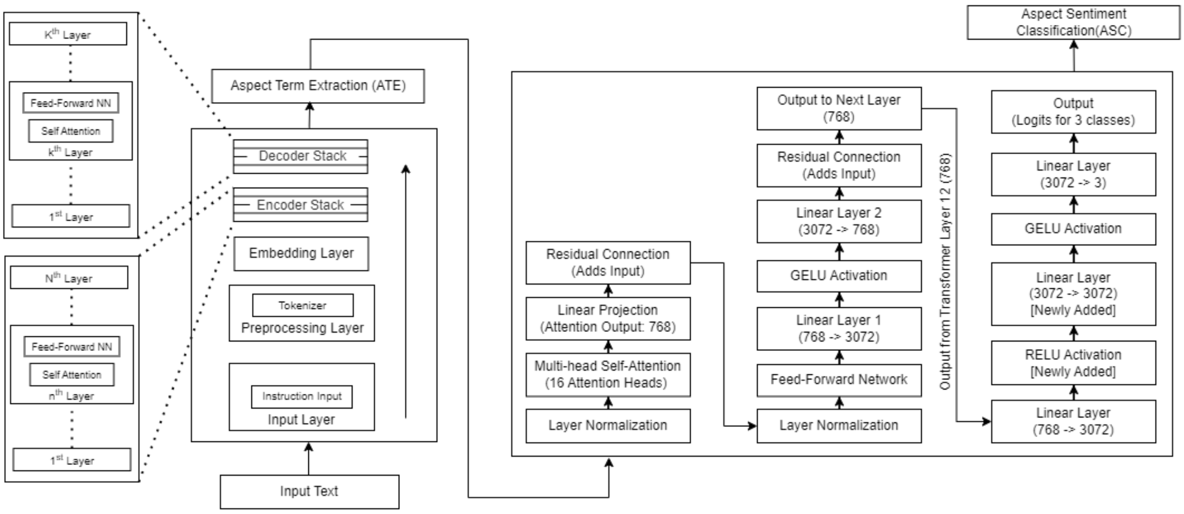


Figure 1: New structure of Instruct-DeBERTa

this categorization process was carried out without training the model on a specific dataset that explicitly links aspects to categories. Instead, it relied entirely on the inherent similarities between embeddings in the vector space, demonstrating the ability of pre-trained embeddings in capturing semantic relationships.

3.2 Improvements to the existing architecture of Instruct-DeBERTa

Under this section, the different changes that we experimented are being discussed for both the aspect extraction and the sentiment polarity model which will eventually increase the performance of the collective hybrid model `Instruct-DeBERTa`. We explored a series of architectural modifications and regularization techniques on the `Instruct-DeBERTa` model to enhance its performance in sentiment analysis tasks. These modifications included adding an extra feed-forward layer, implementing additional regularization methods, and adjusting the number of attention heads. The changes were tested for both the `DeBERTa-V3-baseabsa-V1` which performs ASC and `InstructABSA` which performs ATE. Several of these interventions resulted in improvements to the model's weighted F1 score, highlighting the potential of fine-tuning and architectural adjustments to optimize models with pre-trained weights for specific NLP tasks. This approach emphasizes the value of achieving meaningful performance gains with minimal retraining, reducing the need for extensive re-training with each architectural change.

3.2.1 Adding a linear layer with ReLU and regularization for ASC enhancements

In our experiment, we utilized the `DeBERTa-V3-base-absa-V1` model for the ASC task. The original model's classifier architecture consisted of a linear layer that projected the output of the transformer layers into a higher-dimensional space, followed by a GELU activation function to introduce non-linearity. This was then followed by a final linear layer that reduced the dimensionality to produce logits corresponding to the three sentiment classes (negative, neutral, positive). To explore potential performance improvements, we modified this architecture by adding an additional feed-forward layer in the classifier. Specifically, we introduced an extra linear layer followed by a ReLU activation function after the first linear layer in the classifier. This additional linear layer, which maintained the same output dimensionality, was inserted to perform further transformations of the feature space. The ReLU activation added another layer of non-linearity, enhancing the model's ability to capture complex patterns. By extending the classifier with this deeper architecture, we aimed to increase the model's capacity for more sophisticated feature representations, potentially leading to more accurate classification decisions. There are also additional theoretical grounds for setting feed-forward layers in a universal approximation theorem. The theorem says a neural network with enough depth and non-linearity can approximate any continuous function, and because it adds one more degree of freedom to the model by being flexible in how it models the decision boundary among classes, this might lead to better generalization.

The weighted F1 score, when measured afterwards, improved slightly for Res-14 and Res-15, while it remained the same for Res-16. In addition to that, this represents a marginal yet critical movement toward effectiveness in classification, reflecting the change in realization. In other words, this leads to another layer, hence making the model more effective in capturing base data distribution and representing that, which finally improves prediction accuracy. This documented increase is quite minor in the F1 score but crucial in noting how it may make the model's architecture important to ensure performance is optimized maximally towards the task. Now, with more fine-grained decision-making, that was due to the added feed-forward layer; it brought just a better fit of the model's predictions to the actual labels. This illustrates potential gains of deviation from the base model for general NLP problems in driving up performance. However, these modifications also come with potential disadvantages. The added layers and parameters increase the model's complexity, which introduces a risk of over-fitting, especially if the training data is not large or diverse enough to justify the increased capacity. Over-fitting can cause the model to learn patterns specific to the training data that do not generalize well to unseen data, potentially undermining the benefits of the added complexity (Aliferis and Simon, 2024).

To address the potential over-fitting introduced by adding an extra linear layer and ReLU activation to our model, we explored various regularization techniques. Realizing that the enhanced model complexity led to over-fitting, we resorted to having L2 (ridge) regularization in the classifier of the model (Ying, 2019). This is a method by which large values of weights are penalized so that the model generalizes better to unseen data and does not become very adapted to any specific parameters. n addition to L2 regularization, we also experimented with adjusting the dropout rate to further mitigate over-fitting. So we validated for dropout rates between 0.1 and 0.5, and in the process for the range, there wasn't much significance in changing the accuracy with no re-training. Based on these observations, we selected a dropout rate of 0.3 as a balanced choice for future use. This rate is intended to provide sufficient regularization without overly compromising the model's ability to learn from the training data.

On the other hand, it is also necessary to recognize the threats related to high dropout. Although dropout contributes to model regularization, too much dropout leads to under-fitting: the model poorly learns because the random exclusion of information is too much during the training procedure. This type of situation may marginally impede the ability of the model to fit the training data properly, primarily if the dataset does not possess enough size or diversity. In the process, our strategy for mitigating over-fitting included the implementation of L2 regularization in concert with careful tuning of the dropout rate. These modifications will create a balance between the improvement of generalization and maintaining the learning capability of the model so that it is resilient for use in the future. By incorporating these regularization techniques, we aim to enhance the strength and suitability of the model for future use to ensure it performs its tasks efficiently without over-fitting on the training data.

3.2.2 Increasing the number of attention heads for ASC enhancements

For sentiment classification, we also explored the impact of varying the number of attention heads in the transformer model architecture on the effectiveness of the classification. Attention heads are a crucial component of the multi-head self-attention mechanism in transformer models. Each attention head operates as an independent set of attention mechanisms that learn to focus on different parts or aspects of the input sequence simultaneously. This allows the model to capture diverse patterns and relationships in the data, which are essential for tasks like sentiment classification where multiple contextual cues contribute to the final classification. The number of attention heads determines how many separate attention distributions the model can learn in parallel. Increasing the number of attention heads allows the model to capture more complex patterns and dependencies in the dataset, as each head can focus on different elements of the input sequence (Nguyen et al., 2022).

3.2.3 Improvements done for the aspect term extraction model

In our study related to the aspect term extraction task, we used the same set of architectural changes and a set of regularization methods as described in the previous section for the transformer model-`InstructABSA`, but with pre-trained weights without fine-tuning. In any case, a similar observation was that none of the changes resulted in substantial improvements in the weighted F1 score of the development set for aspect term extraction.

The far less varied F1 score values suggest that the aspect term extraction task may be more sensitive to model architecture and applied regularization techniques than sentiment classification. Moreover, it does not show further improvements in performances due to these modifications, which might indicate that the intrinsic characteristics of aspect term extraction benefited less from the applied changes than what was the case for sentiment analysis tasks. This is likely because of the specialty of the aspect term extraction task itself, which may rely far more on the other dimensions of model performance, or require much more architectural change and regularization than afforded by the experiments.

3.2.4 Integrating the combined model

In the final stage, the enhanced `DeBERTa-V3-baseabsa-V1` ASC model, in which modifications were introduced such as adding a linear layer with ReLU activation with regularization methods and changing attention heads, was combined with the `InstructABSA` ATE model to make the improved version of the combined hybrid model, `Instruct-DeBERTa`. This was supposed to integrate both models' benefits and, as such, integrate their capabilities into one package for comprehensive aspect-based sentiment analysis.

4 Results

Following few key changes to the model, such as, adding an extra linear layer, ReLU, applying regularization methods, and tuning attention head settings, we observed improvements on multiple datasets. These changes were for enhancing the capability of the model to learn complex patterns while retaining its generalization power on previously unseen data. In the following sections, we present a thorough discussion of weighted F1 scores discussing various gains witnessed for the datasets, Res-14, Res-15, and Res-16.

4.1 For integrated aspect categorization

To provide more understanding of the relationships between aspect terms and their categories, we visualized the embeddings using t-SNE for dimensionality reduction and Voronoi diagrams. t-SNE (t-Distributed Stochastic Neighbor Embedding) is a non-linear dimensionality reduction technique that projects high-dimensional data into a 2D or 3D space while preserving the local structure of the data. The embeddings of the aspect terms and categories were reduced from their original high-dimensional space of 768 dimensions to 2D for visualization purposes.

The below cost function is optimized according to the t-SNE algorithm, the function measures the divergence between the probability distributions of the pairwise similarities in the original and target-dimensional spaces:

$$C = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{4}$$

Where:

- P_{ij} is the joint probability that points i and j are neighbors in the high-dimensional space.
- Q_{ij} is the joint probability in the low-dimensional space.

By minimizing this cost function, t-SNE ensures that similar points in the high-dimensional space remain close in the 2D projection. The 2D embeddings of the categories and aspects were then used to generate a Voronoi diagram. A Voronoi diagram partitions the space into regions based on the distance to a set of pre-defined points, known as Voronoi sites.

Mathematically, the Voronoi region V_i associated with a category i is defined as:

$$V_i = \{\mathbf{x} \in \mathbf{R}^2 \mid \|\mathbf{x} - \mathbf{e}_i\| \leq \|\mathbf{x} - \mathbf{e}_j\| \text{ for all } j \neq i\} \tag{5}$$

Where:

- $\mathbf{e}_i$ is the 2D embedding of category i.
- $\|\mathbf{x} - \mathbf{e}_i\|$ is the Euclidean distance between any point $\mathbf{x}$ and the embedding $\mathbf{e}_i$.

The Voronoi diagram as in Figure 2 provides a clear visualization of how each aspect term (projected into the same 2D space) relates to the predefined categories. The regions help in understanding which categories dominate specific areas of the embedding space, and how close or distant different aspects are from each other and their respective categories.

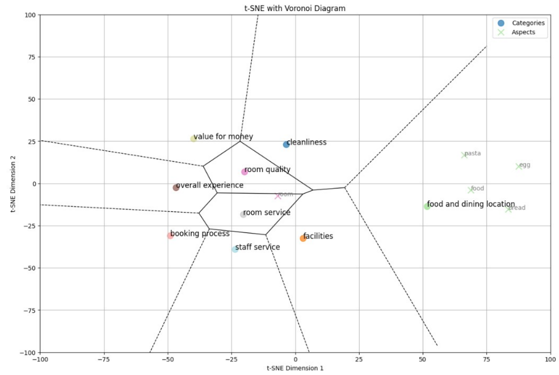


Figure 2: Voronoi diagram to visualize the aspect categories

Under category separation, we mainly focused on the hospitality domain. The pre-defined categories that we used here were cleanliness, facilities, food and dining, booking process, overall experience, room quality, room service, value for money and staff service. We then manually checked the accuracy of the category separation for 100 reviews which were publicly available in the internet, in which we obtained an accuracy of 85%. In the future, we hope to build our own data set for category separation to formally observe the accuracy levels.

4.2 Results after architectural improvements for Instruct-DeBERTa

This section highlights the enhancements in weighted F1 resulting from the changes discussed in the methodology section. The modifications are carefully tested to ascertain their impact on model performance with respect to ABSA in the hospitality industry. A comparison of F1 scores between the enhanced and standard models clearly underlines the efficiency of the revised methodology. The upgraded model gave better overall performance proving that its performance enhancement was prominent, and thus the precision and generalization capability are significantly higher.

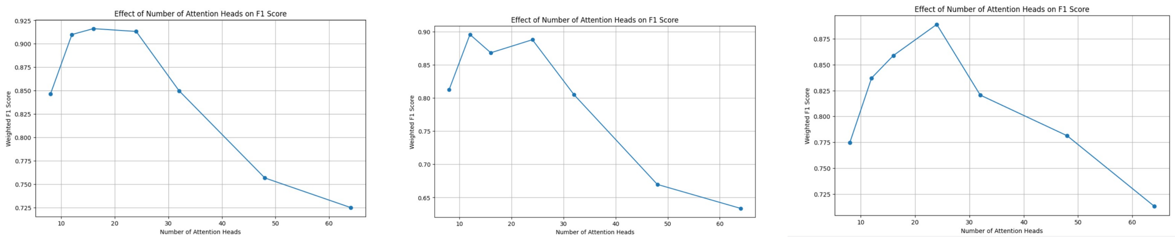
4.2.1 After adding a linear layer with ReLU and regularization

Initially, the weighted F1 scores achieved for Res-14, Res-15, and Res-16 were 90.94%, 89.55%, and 83.71% respectively as in Table 1. After adding an extra linear layer followed by a ReLU activation function after the first linear layer in the classifier, it was observed that the F1 scores for Res-14 and Res-15 improved to 90.99% and 89.56% respectively while the F1 score for Res-16 remained the same. Changing the dropout rates and applying L2 regularization for the classifier did not result any change in the F1 scores but they were added to the model to overcome over-fitting as discussed in the methodology section.

4.2.2 After Increasing the number of attention heads

We tested the model with various numbers of attention heads, starting from 8, 12, 16, 24, 32, 48, and 64 heads, respectively. The default value was 12 attention heads, which aligns with the model's hidden state size of 768. In transformer models, the number of attention heads must be a divisor of the hidden size to ensure that each head receives an equal portion of the hidden representation. This is why divisors of 768 were chosen for the experiment—ensuring that the hidden state size could be evenly split across the attention heads without causing errors during processing. The F1 scores were calculated by varying the number of attention heads for all three data sets as in Figure 3.

For Res-14, the resulting weighted F1 scores were 0.8462, 0.9099, 0.9162, 0.9131, 0.8497, 0.7565, and 0.7249 for attention heads 8, 12, 16, 24, 32, 48, and 64 respectively. These results indicate that increasing the number of attention heads initially enhances the model's ability to learn and generalize by capturing a wide range of attention patterns. Specifically, with 12,16, and 24 attention heads, the model achieved the highest F1 scores of 0.9099, 0.9162, and 0.9131 respectively. This suggests that at these levels, the model achieves an optimal balance, providing enough parallel attention distributions to capture complex data dependencies without overwhelming its learning capacity. However, as the number of attention heads increased above 16, the performance began to decline. The F1 scores dropped significantly as the attention heads were increased to 32, 48, and 64. The reason for the decline is due to the over-parameterization of the model. As the attention heads increase, the model will begin to overfit for the training data and lose its ability to generalize for unseen data (Voita et al., 2019). Additionally, when the model is made complex with too many attention heads, each head may receive fewer computational resources, leading to weaker attention distributions and less effective learning (Michel et al., 2019). Our findings indicated that for the ASC task of Res-14, 16 attention heads provided the best performance, resulting in the highest F1 score of 0.9162. This was achieved by using the same pre-trained weights ini-



(a) Res-14 (b) Res-15 (c) Res-16

Figure 3: Variation in F1 score with an increase in the number of attention heads

tially trained with 12 attention heads, demonstrating that careful tuning of the model architecture can lead to significant performance improvements without the need for extensive retraining. In addition to that for the Res-16 data set the same phenomenon was observed, where unlike in Res-14 the peak F1 score was achieved at 24 attention heads while for Res-15 the peak was observed at the default 12 attention heads. To balance these variations and optimize performance across different datasets, we selected the mid-value of 16 attention heads for our final model.

4.2.3 Joint task F1 scores for improved Instruct-DeBERTa

The performance of the integrated model was quantified for F1 scores on the joint task, hence bearing insights into important perspectives about the improvement in overall performance achieved by such integration. Here joint Task F1 Scores refers to the performance metric calculated for the entire pair of aspect and sentiment as a combined task, rather than evaluating them separately. In this context, the model's performance is assessed based on its ability to correctly identify both the aspect term and its corresponding sentiment in a sentence. This means that the F1 score reflects the model's accuracy for not just extracting the correct aspect but also assigning the correct sentiment to the respective aspect. As in Table 1 for the single task of the combined model, F1 score values remained the same across the three data sets for the ATE task since no architectural changes were made. However, the ASC F1 Scores increased for Res-14 and Res-16 significantly with the changes. The ASC F1 value remains the same for Res-15 since it peaks at 12 attention heads and we have used 16 to suit all the data sets as a whole. Furthermore, as observed in Table 1 the joint task F1 score for Res-14 also improved by 1.14%. The joint task F1 scores were not previously calculated for the other two data sets, hence we calculated them and included in Table 1. In addition to those, we checked the F1 score for the Lap-14 dataset as well. It also improved from 80.94% (Jayakody et al., 2024b) to 80.97%. The improved version shows promising results across multiple domains, demonstrating that it works well for other domains too. However, the model can be further customized to optimize its performance when the domain changes, allowing for better adaptation and fine-tuning to specific domain characteristics.

5 Conclusion

In this work, we aimed at improving the `Instruct-DeBERTa` model by focusing its base models individually. The improvements added were a linear layer followed by ReLU activation, incorporation of regularization, optimization of attention heads, and adding an aspect category extraction capability. Importantly, this was done without retraining the model; thus, it demonstrates our approach toward enhancing the model's performance without losing those strengths it previously demonstrated. These strategic adjustments indeed caused significant enhancement in the weighted F1 scores across the datasets, especially in the hospitality domain. The model was further augmented by incorporating the function of aspect category extraction that allowed the model to go beyond just the identification of aspects and sentiments but instead classify aspects effectively. Improvement within the `Instruct-DeBERTa` hybrid model concretizes a path toward realizing significant accuracy gain on domain-specific sentiment analysis applications. Further optimizations can be explored in future studies and this method can be applied to other domains for the expansion of applicability and effectiveness as well.

References

Constantin Aliferis and Gyorgy Simon. 2024. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and ai. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, pages 477–524. Springer.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *NeurIPS*, volume 36.

Dineth Jayakody, Koshila Isuranda, AVA Malkith, Nisansa de Silva, Sachintha Rajith Ponnamperuma, GGN Sandamali, and KLK Sudheera. 2024a. Aspect-based sentiment analysis techniques: A comparative study. *arXiv preprint arXiv:2407.02834*.

Dineth Jayakody, A V A Malkith, Koshila Isuranda, Vishal Thenuwara, Nisansa de Silva, Sachintha Rajith Ponnamperuma, G G N Sandamali, and K L K Sudheera. 2024b. Instruct-DeBERTa: A Hybrid Approach for Aspect-based Sentiment Analysis on Textual Reviews. *Preprint*, arXiv:2408.13202.

Sahan Jayasinghe, Lakith Rambukkanage, Ashan Silva, Nisansa de Silva, and Amal Shehan Perera. 2021. Party-based Sentiment Analysis Pipeline for the Legal Domain. In *2021 21st International Conference on Advances in ICT for Emerging Regions (ICter)*, pages 171–176. IEEE.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Chanika Ruchini Mudalige, Dilini Karunarathna, Isanka Rajapaksha, Nisansa de Silva, Gathika Ratnayaka, Amal Shehan Perera, and Ramesh Pathirana. 2020. SigmaLaw-ABSA: Dataset for Aspect-Based Sentiment Analysis in Legal Opinion Texts. In *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, pages 488–493. IEEE.

Tan Nguyen, Tam Nguyen, Hai Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Khuong Duy Nguyen, Nhat Ho, and Stanley Osher. 2022. Improving transformer with an admixture of attention heads. *Advances in neural information processing systems*, 35:27937–27952.

Isanka Rajapaksha, Chanika Ruchini Mudalige, Dilini Karunarathna, Nisansa de Silva, Amal Shehan Perera, and Gathika Ratnayaka. 2021. Sigmalaw PBSA-A Deep Learning Model for Aspect-Based Sentiment Analysis for the Legal Domain. In *International Conference on Database and Expert Systems Applications*, pages 125–137. Springer.

Isanka Rajapaksha, Chanika Ruchini Mudalige, Dilini Karunarathna, Nisansa de Silva, Gathika Rathnayaka, and Amal Shehan Perera. 2020. Rule-Based Approach for Party-Based Sentiment Analysis in Legal Opinion Texts. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 284–285. IEEE.

Isanka Rajapaksha, Chanika Ruchini Mudalige, Dilini Karunarathna, Nisansa de Silva, Gathika Ratnayaka, and Amal Shehan Perera. 2022. Sigmalaw PBSA-A Deep Learning Approach for Aspect-Based Sentiment Analysis in Legal Opinion Texts. *J. Data Intell.*, 3(1):101–115.

Chamodi Samarawickrama, Melonie de Almeida, Nisansa de Silva, Gathika Ratnayaka, and Amal Shehan Perera. 2022. Legal Party Extraction from Legal Opinion Texts Using Recurrent Deep Neural Networks. *J. Data Intell.*, 3(3):350–365.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 720–736, Mexico City, Mexico. Association for Computational Linguistics.

Aivin V. Solatorio. 2024. GISTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning. *arXiv preprint arXiv:2402.16829*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Heng Yang and Ke Li. 2024. Modeling aspect sentiment coherency via local sentiment aggregation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 182–195, St. Julian's, Malta. Association for Computational Linguistics.

Heng Yang, Biqing Zeng, Mayi Xu, and Tianxing Wang. 2021. Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning. *arXiv preprint arXiv:2110.08604*.

Heng Yang, Chen Zhang, and Ke Li. 2023. Pyabsa: A modularized framework for reproducible aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 5117–5122. ACM.

Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.