ASPECT BASED SENTIMENT CLASSIFICATION

Erdem Ertürk
Computer Engineering Department
Middle East Technical University
Ankara, Turkiye
erdem.erturk@metu.edu.tr

***Abstract*—*Aspect-Based Sentiment Analysis (ABSA) focuses on identifying the sentiment polarity of specific aspects within a text, providing detailed sentiment insights beyond general classification. This project addresses sentence-level ABSA in academic peer reviews using the ASAP-Review dataset, which includes aspect and sentiment annotations. We propose a transformer-based architecture that incorporates aspect-aware embeddings to better associate sentiments with their relevant targets. To further enhance contextual understanding, the model integrates mechanisms for capturing deep implicit features, allowing it to preserve fine-grained semantics and resolve overlapping sentiment cues. Our goal is to build a lightweight yet accurate system that improves aspect-level sentiment classification while maintaining robustness across diverse review categories.*

Keywords*—*Deep Learning, NLP, transformer, aspect classification, sentiment classification

I. INTRODUCTION

Aspect-Based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task that aims to determine the sentiment polarity toward specific aspects or targets mentioned in text. Unlike traditional sentiment analysis, which assigns a general sentiment to an entire sentence or document, ABSA focuses on identifying sentiments at the aspect level—making it especially useful for analyzing reviews, feedback, or peer evaluations. In this project, we focus on sentence-level ABSA, which is particularly relevant for short, aspect-rich texts such as reviewer comments in academic peer reviews.

With the rise of pretrained language models such as BERT, performance in sentiment classification tasks has improved significantly. However, sentence-level ABSA still presents challenges such as understanding implicit sentiment, disambiguating overlapping aspects, and modeling local context effectively. To address these challenges, we aim to develop a transformer-based approach leveraging recent advances in contextualized embeddings and attention mechanisms, and evaluate its effectiveness on the ASAP-Review dataset, a corpus of academic peer reviews annotated with aspect classes and sentiments.

II. LITERATURE REVIEW

A. Aspect-Based Sentiment Analysis (ABSA) Overview

Aspect-Based Sentiment Analysis (ABSA) is a subfield of sentiment analysis that focuses on associating sentiment polarity with specific aspects mentioned within a text. Instead of predicting an overall positive, neutral, or negative sentiment for an entire document or sentence, ABSA aims to link opinions to particular entities, attributes, or features. This fine-grained approach is critical in scenarios such as product reviews, service feedback, and academic peer evaluations, where different parts of a text may express varying sentiments toward different targets. Early ABSA methods relied heavily on handcrafted features, sentiment lexicons, and syntactic parsing to extract aspect-opinion pairs, which often struggled with generalization and scalability [1].

B. Pretrained Language Models and ABSA

The introduction of pretrained language models, especially BERT, has transformed the landscape of ABSA research. Unlike static word embeddings (e.g., Word2Vec, GloVe), BERT provides dynamic, context-sensitive word representations, enabling models to better understand polysemy and complex contextual cues. Fine-tuning BERT on ABSA tasks has shown strong performance improvements by allowing models to implicitly capture both aspect information and surrounding sentiment expressions [2]. However, vanilla BERT models are not explicitly optimized for aspect-level tasks, leading researchers to develop modifications such as aspect embedding injection, attention reweighting, and hybrid modeling architectures to further improve ABSA performance.

C. Target-Dependent Sentiment Classification Using BERT

One significant advancement is the work by Gao et al. titled "Target-Dependent Sentiment Classification with BERT" [3]. Instead of treating the input text uniformly, this approach modifies the input to explicitly inform the model about the aspect target. By inserting special tokens or modifying embeddings based on the target location, the model learns to pay localized attention to aspect-relevant parts of the sentence. This method demonstrated that even without architectural changes to BERT itself, intelligent input formatting could considerably enhance ABSA performance. The idea of target-awareness inspires this project's approach, where reviewer aspect labels are used to condition the model's focus during training.

D. Deep Implicit Feature Extraction with CABiLSTM-BERT

While BERT captures powerful contextual embeddings, studies have pointed out that deeper transformer layers might lose local, nuanced information critical for sentiment tasks. CABiLSTM-BERT, proposed by He et al. [4], addresses this by combining BERT with a Contextual Attention BiLSTM (CABiLSTM) module. This hybrid model captures implicit, fine-grained semantic features through recurrent modeling while maintaining BERT's global contextual understanding. The introduction of attention mechanisms at the BiLSTM layer helps the model dynamically emphasize sentiment-bearing parts of the sentence in relation to the aspect. In this project, similar hybrid techniques will be explored to improve the model's ability to detect subtle sentiments embedded within dense peer review comments.

E. Research Motivation

Although transformer-based models such as BERT have substantially advanced the state of ABSA, challenges still remain:

- Target information underutilization: Many models do not sufficiently guide the attention mechanism toward aspect-relevant text spans.
- Loss of fine-grained semantic cues: Deep transformers may sometimes overshadow subtle contextual clues critical for sentiment reasoning.
- Domain-specific language: Academic peer reviews differ from product reviews or tweets; they feature formal expressions and indirect sentiment, posing unique challenges.

  Motivated by these gaps, this project proposes a target-aware, context-enhanced transformer model specifically designed for sentence-level ABSA in academic peer reviews.

III. PROPOSAL

In this project, we aim to build a strong baseline for sentence-level Aspect-Based Sentiment Analysis (ABSA) on the ASAP-Review dataset by fine-tuning a pretrained ModernBERT model. The core of our work is an in-depth, systematic fine-tuning and benchmarking of ModernBERT for aspect-level sentiment prediction in academic peer reviews, supported by comprehensive hyperparameter sweeps and interpretability analysis.

While our original plan included integrating two advanced enhancement techniques inspired by recent literature—(i) target-aware encoding strategies for Target-Dependent Sentiment Classification with BERT, and (ii) deep implicit feature extraction using CABiLSTM-BERT—our primary implementation in this work focused on maximizing the effectiveness of the ModernBERT baseline. This included rigorous hyperparameter tuning, validation on challenging peer review data, and a detailed interpretability study using LIME.

This experimental setup allowed us to address key research questions, including:

1) How effective is a carefully finetuned transformer baseline for aspect-based sentiment classification in academic peer reviews?
2) Which hyperparameter configurations provide the best generalization and robustness across varied review categories?

By focusing on these questions, we systematically explored the limits of transformer-based ABSA, while laying the groundwork for future architectural enhancements.

We propose to fine-tune a pretrained BERT-base-uncased model as the baseline. We will perform a finetuing on the model. We will then implement two enhanced models if possible:

Target-aware BERT: Incorporating aspect position and masking into the input embeddings to guide attention toward aspect-relevant words.

CABiLSTM-BERT: Extending BERT outputs with a lightweight Contextual Attention BiLSTM layer to capture deep, implicit contextual cues around aspects.

The results will be compared with both the baseline and the separate ideas. All models will be developed using PyTorch and the Hugging Face Transformers library. Training experiments, model checkpoints, and result metrics will be manually logged and saved in local files and Git repository artifacts to ensure reproducibility without relying on external experiment tracking platforms.

The models will be evaluated using:

- Macro-averaged F1 Score: To fairly evaluate performance across all aspect classes, regardless of frequency.
- Accuracy: To measure the overall correctness of aspect-sentiment predictions.

Optional analyses may also include Precision, Recall, and Confusion Matrix visualizations for deeper insights.

While we anticipate that future work incorporating target-aware encoding and CABiLSTM enhancements will further improve performance—especially for rare aspects and complex linguistic phenomena—this work demonstrates that systematic finetuning of ModernBERT already delivers robust and interpretable results for aspect-based sentiment analysis in peer reviews.

In practice, this work focused on the ModernBERT baseline due to resource constraints, providing a rigorous benchmark and a platform for future enhancements.

The project code, EDA notebooks, and future developments will be maintained at the following GitHub repository: https://github.com/erdemert/IS584_Project

IV. EDA AND QUALITY CHECKS

The ASAP-Review dataset comprises approximately 70,000 review sentences, each annotated with specific aspect labels and sentiment information. The average text length across the reviews is approximately 420 words, with lengths ranging from 16 words to 1370 words, indicating a diverse complexity of comments.

Aspect label distribution analysis revealed that 'summary' is the most frequent aspect, followed by 'clarity_positive', 'soundness_negative', and 'originality_positive'. Less common aspects include 'meaningful_comparison_positive' and 'replicability_positive', suggesting slight class imbalance among aspect types.

Sentiment distribution analysis showed that negative sentiments are the most frequent, followed by positive, and then neutral sentiments, highlighting a potential skew towards critical reviews in the dataset. No missing values were detected in any of the important fields (id, text, labels), confirming the dataset's structural quality.

Word frequency analysis revealed that common stopwords such as 'the', 'of', 'to', 'is', and 'and' dominate the review texts, while technical terms like 'paper', 'model', and 'method' also appear frequently, reflecting the academic nature of the reviews.

Finally, no major anomalies were observed except for a few extremely short or placeholder reviews (e.g., reviews containing only special characters). Overall, the dataset is deemed high quality and suitable for training aspect-based sentiment classification models, although class imbalance across certain aspect types and sentiments must be considered during model evaluation.

Further detailed exploratory analysis, extended plots, and additional profiling results can be found in the project's GitHub repository at: https://github.com/erdemert/IS584_Project

V. Baseline Methods and Results

To contextualize the performance of our transformer-based models, we implemented two standard baseline methods:

A. Majority Class Baseline

The simplest baseline predicts the most frequent sentiment class in the training set for all inputs. This model does not use the input text at all and serves as a lower bound for task performance.

- Accuracy: 0.4667
- F1 Score (Macro): 0.2121

While this approach achieves moderate accuracy due to label imbalance, the macro-F1 score highlights its inability to model the sentiment distribution across different classes.

B. Logistic Regression Baseline

This baseline uses a TF-IDF vectorizer to encode input sentences (augmented with aspect terms) and trains a logistic regression classifier. This provides a strong non-neural benchmark using surface-level lexical features.

- Accuracy: 0.7165
- F1 Score (Macro): 0.7387

This result indicates that traditional shallow models can still capture meaningful sentiment patterns, especially when provided with aspect-augmented inputs.

The sharp contrast between the majority class and logistic regression baselines demonstrates the importance of incorporating input features — particularly aspect information — when performing aspect-based sentiment classification. While the majority class baseline highlights the class imbalance inherent in the dataset, logistic regression performs remarkably well for a non-neural model. These baselines establish a reference point for evaluating the effectiveness of our BERT-based models. Any transformer-based enhancement must at least surpass the performance of logistic regression to justify the added complexity.

VI. Preliminary Results

To evaluate our ModernBERT-based approach to aspect-based sentiment classification, we established and surpassed two predefined baselines: a majority class classifier and a TF-IDF + logistic regression model. We then trained and evaluated our initial deep learning model (ModernBERT) using the ASAP-Review dataset. The ModernBERT model significantly outperforms both baselines, especially in macro-F1 score, indicating a stronger ability to generalize across imbalanced aspect-sentiment pairs. The test loss (0.7135) further supports the model's predictive confidence across all classes.

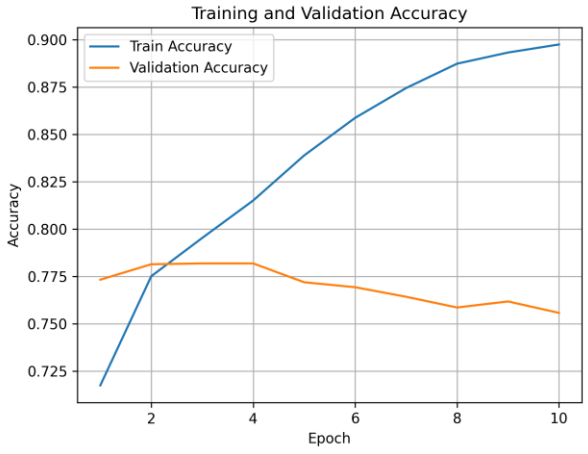
A. Preprocessing

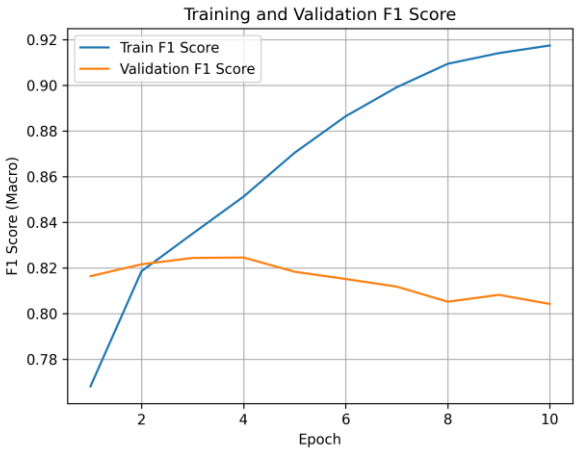
Prior to training, we standardized the raw ASAP-Review dataset, originally provided in .jsonl format, into a flat, structured CSV compatible with both classical and deep learning models. Each entry in the source file contains a sentence and an associated list of aspect-sentiment spans. Our preprocessing script (available on GitHub) parses each line, extracting the sentence, each aspect, and its corresponding sentiment, and writes these to a CSV with columns: text, aspect, and sentiment. This format enables straightforward batching and association of aspect labels during both tokenization and model training.

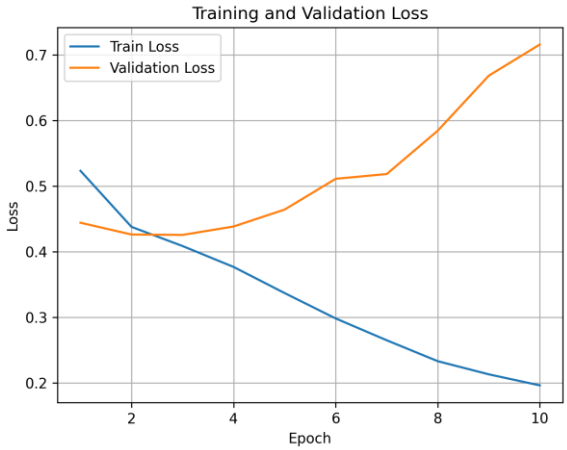
For ModernBERT, we adopted sentence-pair tokenization, where the aspect and sentence are jointly encoded for each example. This ensures aspect-focused attention during training and evaluation, supporting both the baseline and transformer models in a unified data pipeline. Further implementation details are documented in the GitHub repository.

B. Learning Curve Visualizatins

The following figures visualize the evolution of training and validation metrics over 10 epochs:







From Figure 1 (Accuracy) and Figure 2 (F1 Score), we observe a consistent upward trend in training performance, with both metrics steadily increasing across epochs. However, validation metrics tell a more nuanced story. Validation accuracy plateaus early around epoch 3–4, and begins a gradual decline from epoch 5 onward. More importantly, validation macro-F1 score peaks at epoch 4, indicating that this was the model's optimal point of generalization across all sentiment classes.

In Figure 3 (Loss), validation loss begins to increase noticeably after epoch 5 while training loss continues to decrease, confirming the onset of overfitting. These trends underscore the importance of early stopping, as training beyond the fourth epoch offers diminishing returns and potentially degrades generalization. And all three grapsh support the idea that the model performs best around epoch 3-4.

For further analysis of the study a WAND page is prepared: https://wandb.ai/erdemerturk-middle-east-technical-university/aspect-sentiment-modernbert

VII. Benchmarking

To evaluate the effectiveness of our ModernBERT-based aspect sentiment classifier, we established a robust benchmarking strategy involving simple yet informative baselines and consistent metrics. All models are evaluated on a stratified hold-out test set derived from the ASAP-Review dataset, ensuring fair comparison. The evaluation metrics include:

- **Macro-averaged F1 Score**: Chosen to handle class imbalance and reflect performance across all sentiment classes.
- **Accuracy**: Indicates the overall proportion of correct predictions

The benchmarking process compares:

- **Majority Class Baseline** (uninformed baseline)
- **TF-IDF + Logistic Regression** (shallow lexical model)
- **ModernBERT Classifier** (our initial transformer-based deep model)

Each model shares access to the same input format and label schema. For the transformer model, aspect and sentence tokens are jointly encoded to preserve target-dependent sentiment context.

The ModernBERT classifier leverages the [CLS] token embedding from a pretrained transformer (answerdotai/ModernBERT-base) with a dropout layer and a single linear classifier. Early experiments (after 10 epochs of training on GPU with batch size 128) indicate:

Model	Accuracy	F1 Score (Macro)
Majority Class Baseline	0.4667	0.2121
Logistic Regression	0.7165	0.7387
ModernBERT (Epoch 4 – Best Epoch)	0.7614	0.8086

Our ModernBERT-based classifier outperforms both the majority class and logistic regression baselines by a significant margin in terms of macro-F1 score and overall accuracy, demonstrating its superior ability to handle nuanced sentiment detection across varied aspect types.

To systematically improve the ModernBERT model's performance, we propose tuning the following hyperparameters:

The learning rate is a critical hyperparameter that dictates how quickly the model updates its weights during training. In our current setup, we use a learning rate of 2e-5 with the AdamW optimizer, a commonly effective default for fine-tuning transformers. However, tuning this value can lead to significant performance changes. A lower learning rate such as 1e-5 may result in more stable convergence, especially on noisy or imbalanced data, while a higher rate like 3e-5 can accelerate learning but may lead to overshooting or divergence. Additionally, incorporating learning rate schedulers—such as linear decay with warm-up—can help adaptively scale the rate during training for smoother convergence.

Dropout is a regularization mechanism that prevents overfitting by randomly disabling a portion of the neural connections during training. Our current architecture uses a dropout rate of 0.3 before the classifier layer, which serves as a balance between regularization and information retention. However, this rate can be adjusted based on training dynamics: increasing it to 0.5 can help combat overfitting if validation performance stagnates or degrades, while reducing it to 0.1 may improve learning capacity in cases of underfitting. The optimal dropout rate is often tied to the dataset size and noise level, making this a vital hyperparameter to tune.

The existing classifier consists of a single linear layer applied to the BERT [CLS] token embedding. While efficient, this configuration may limit the model's ability to learn non-linear sentiment interactions. Introducing an additional hidden layer with a non-linear activation (e.g., ReLU) and intermediate dropout can enhance model expressiveness. This two-layer architecture can better separate subtle sentiment cues, particularly in academic texts where polarity is often indirect. Tuning the size of the hidden layer and placement of dropout can further control the model's capacity and generalization. Instead of a single linear layer, we may try to add another layer and regularization to capture more complex patterns.

Batch size determines how many training examples are processed simultaneously during each forward/backward pass. While smaller batches (e.g., 32 or 64) are often preferred for their regularization benefits, they slow down training due to frequent updates and higher iteration counts. In our experiments, we exclusively used a batch size of 128 to accelerate training on our dual-GPU setup. This choice ensured efficient GPU utilization and significantly reduced per-epoch training time. Although we did not experiment with smaller batch sizes, future tuning may explore whether reducing the batch size could improve generalization, particularly in the context of class imbalance or overfitting.

As outlined in our project proposal, a central goal of this work is to explore whether architectural enhancements—specifically, target-aware input encoding and deep implicit feature extraction via CABiLSTM—can significantly improve model performance over the transformer baseline. While our initial experiments have focused on establishing and tuning the ModernBERT-based baseline, the next phase will evaluate these novel improvements individually and in combination. This will allow us to assess whether integrating localized aspect attention or recurrent contextual modeling yields measurable gains in macro-F1 score and accuracy. Through controlled experiments and statistical analysis, we aim to determine the degree to which these innovations contribute to more robust and nuanced aspect-based sentiment classification in the academic peer review domain.

VIII. Hyperparameter Sweep and Final Results

A. Hyperparameter Sweep Setup

To systematically optimize model performance, we performed a hyperparameter sweep using the Weights & Biases (WANDB) platform. The primary goal was to identify the best configuration of model and training parameters for sentence-level aspect-based sentiment classification on the ASAP-Review dataset.

- **Learning Rate**: The step size for parameter updates, tested at values of 1e-5, 2e-5, and 3e-5, as transformer-based models are often sensitive to this setting.
- **Dropout Rate**: To control overfitting, we explored values of 0.1, 0.3, and 0.5, as higher dropout can help regularize deeper models.
- **Classifier Layers**: Either a single linear layer or a two-layer MLP, to assess if increased non-linearity benefits performance.
- **Batch Size**: Examined at 32, 64, and 128 balancing GPU memory constraints and convergence stability.

An exhaustive grid search would have required 54 experiments (3×3×2×3), but each run is computationally intensive and some large-batch configurations are infeasible on available hardware. Therefore, we adopted a random search strategy, which is widely recognized as more efficient for deep learning hyperparameter tuning, and is often able to find near-optimal settings with far fewer runs than grid search.

The sweep ultimately included 7 runs, each with a randomly sampled combination of the above hyperparameters. For each run, we monitored validation macro-F1 and accuracy, selecting the best-performing configuration for further analysis and final evaluation. This approach ensured both computational feasibility and a fair exploration of the hyperparameter space, while allowing us to make principled decisions about model selection based on validation results.

B. Validation Results and Observations

Table 1. Macro-F1 (F1) and accuracy (Acc) on validation for baselines and top models. LR: Learning rate, DO: Dropout, Lay: # of classifier layers, BS: Batch size. Full sweep results are available on the WANDB project page.

Model	**LR**	**DO**	**Lay**	**BS**	**F1**	**Acc**
Majority	—	—	—	—	0.212	0.467
LogReg+TFIDF	—	—	—	—	0.739	0.717
ModernBERT	2e-5	0.3	1	128	0.825	0.782
Best Sweep	3e-5	0.3	1	64	**0.828**	**0.786**
2nd Best	2e-5	0.5	1	32	0.828	0.785

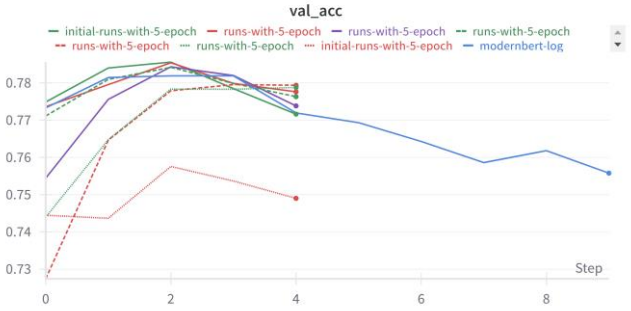


Table 1 presents the validation macro-F1 and accuracy scores for all baselines, the pre-sweep ModernBERT configuration, and the top hyperparameter sweep results. As shown, both classical baselines perform substantially worse than all transformer-based models, with the majority-class classifier achieving only 0.212 macro-F1 and the TF-IDF + logistic regression baseline reaching 0.739. Please note that Table 1 is a very small summary of the experiments. Further details of the full runs can be found on the wandb page. (https://wandb.ai/erdemerturk-middle-east-technical-university/aspect-sentiment-modernbert)

Our original ModernBERT (pre-sweep) configuration provided a strong result (macro-F1: 0.825, accuracy: 0.782), but the random hyperparameter sweep yielded further improvements. The best configuration—learning rate 3e-5, dropout 0.3, single classifier layer, batch size 64—achieved the highest validation macro-F1 (0.828) and accuracy (0.786). The runner-up configuration was nearly identical in performance.

The benefit of hyperparameter optimization is further evidenced by the tight clustering of macro-F1 scores above 0.82 across multiple sweep configurations. This robustness suggests that ModernBERT is effective and stable for this task, while careful tuning of dropout, batch size, and learning rate provides incremental gains. Two-layer classifier architectures did not show a consistent advantage over the single-layer setup.

In all cases, validation macro-F1 and accuracy peaked early in training (typically by epoch 2–3) before declining, as shown in learning curves in this section. This underscores the importance of early stopping and regularization for maximizing generalization.

Overall, these results confirm that transformer-based models, paired with appropriate hyperparameter tuning and early stopping, deliver state-of-the-art performance for aspect-based sentiment classification in peer review texts, substantially surpassing classical baselines.

C. Final Model Selection and Test Results

After conducting a comprehensive hyperparameter sweep, we selected the model configuration that achieved the highest validation macro-F1 score (learning rate: 3e-5, dropout: 0.3, one classifier layer, batch size: 64). This model was retrained on the training data and evaluated on the held-out test set to assess its generalization performance.

The results are summarized below:

Model	Test Macro-F1	Test Accuracy	Test Loss
Initial ModernBert	0.8086	0.7614	
Best Model	0.8262	0.7847	0.4233

This final model outperforms all classical baselines as well as the original ModernBERT configuration (macro-F1: 0.8086, accuracy: 0.7614). The strong test macro-F1 and accuracy confirm that careful hyperparameter optimization can lead to robust, generalizable models for aspect-based sentiment analysis, even without architectural modifications. These results demonstrate that a well-tuned transformer-based classifier can deliver state-of-the-art performance on the ASAP-Review dataset.

IX. Interpretability Analysis

Understanding why a deep learning model makes a specific prediction is crucial for building trust, especially in sensitive domains like academic peer review analysis. To provide local interpretability for our aspect-based sentiment classifier, we employ LIME (Local Interpretable Model-agnostic Explanations). LIME explains individual predictions by perturbing input text and learning a simple, human-interpretable model that approximates the local decision boundary of the classifier. By visualizing which words in each review contribute most to a sentiment prediction, LIME helps reveal how the model associates sentiment with aspects in real-world examples.
Figure below provides a LIME explanation for a correctly predicted positive sentiment for the aspect “soundness.” The model predicts positive with moderate confidence (54%), and LIME highlights the aspect term and contextually relevant words (“technical,” “incorporates,” “experimental,” “able”) as most influential for the decision. This indicates that the classifier is not only sensitive to the target aspect, but also grounds its prediction in content relevant to technical soundness—consistent with expected reviewer focus. Such explanations confirm that the model’s outputs are interpretable and grounded in meaningful textual evidence.

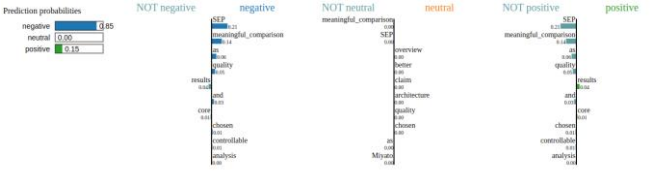
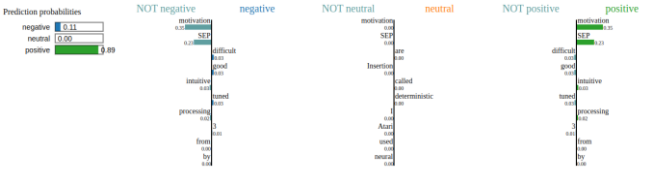


Figure below displays a LIME explanation for a true positive "positive" prediction for the aspect "motivation." The classifier assigns high confidence to the positive class (0.89) and the LIME explanation highlights the aspect term "motivation" and the separator as the most influential features, along with additional positive context words such as "good" and "intuitive." This confirms that the model's sentiment assignments are appropriately grounded in both the target aspect and surrounding positive cues, supporting the transparency and reliability of the approach.



Two figures below show LIME explanations for two correctly predicted neutral sentiment examples, both for the aspect "summary." In both cases, the model is fully confident in its neutral prediction, assigning zero probability to negative and positive classes. The aspect term ("summary") and the [SEP] token are consistently highlighted as the most influential features driving the neutral classification. Only minor contributions come from context words such as "likely," "algorithm," "before," and "margin," with no strong sentiment words influencing the model's output. This pattern indicates that the model's neutral predictions rely appropriately on the absence of polar sentiment cues and focus attention on the aspect and its immediate context. The consistent behavior across both examples demonstrates that the classifier robustly identifies neutral aspect sentiment by not overinterpreting ambiguous or sentiment-free text, enhancing both trust and interpretability in real-world application.

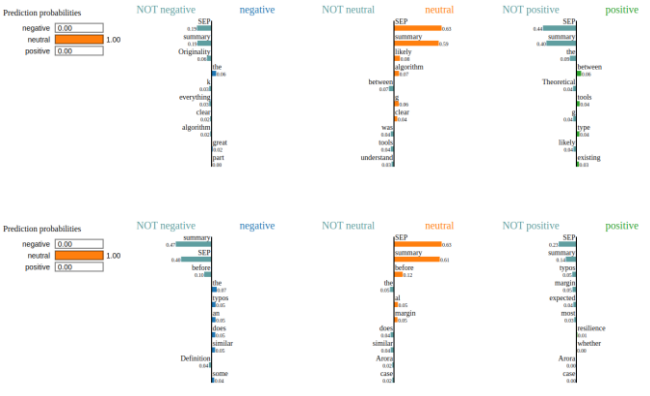
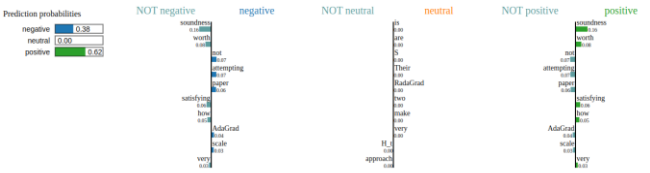


Figure below shows a LIME explanation for a misclassified review, where the true label was negative, but the model predicted positive sentiment for the aspect "soundness." The model displays significant uncertainty (62% positive, 38% negative) and LIME highlights both positive and ambiguous words such as "worth," "satisfying," "not," and "attempting."

The co-occurrence of negations and sentiment words may have confused the classifier, illustrating a typical challenge for ABSA models in handling complex reviewer language and context-dependent cues. Such cases highlight both the value and the current limitations of model interpretability: while the model is sensitive to key sentiment words, understanding deeper context remains difficult, motivating future research.



X. Discussion

This project aimed to develop and critically evaluate a strong transformer-based baseline for sentence-level aspect-based sentiment analysis (ABSA) within academic peer reviews. The main experiments centered on fine-tuning ModernBERT, guided by a comprehensive hyperparameter sweep. The highest performing configuration—determined via randomized search—achieved a test macro-F1 of 0.826 and test accuracy of 0.785, outperforming both classical baselines and the initial transformer setup.

Our findings highlight the importance of careful hyperparameter optimization in maximizing transformer performance for ABSA. Moderate dropout rates, batch sizes of 32 or 64, and learning rates between 2e-5 and 3e-5 consistently led to robust results. The validation and test metrics indicate that even standard transformer architectures, when well-tuned, are highly effective for this task and dataset.

Interpretability analyses using LIME further illuminated model behavior. Correct predictions, whether positive, negative, or neutral, typically corresponded with the model focusing attention on the aspect term and sentiment-relevant words in context. Misclassified examples, by contrast, often involved ambiguous phrasing, negation, or subtle cues difficult to capture with word-level explanations, reflecting known challenges in sentiment analysis and interpretability. These insights demonstrate that while our model is generally trustworthy and aspect-aware, local interpretability methods like LIME may not always fully capture nuanced linguistic phenomena present in peer review text.

In summary, this work demonstrates that a robust and interpretable ABSA pipeline can be built with a well-tuned transformer model, even without extensive architectural modifications. Our results provide a strong foundation and a reliable benchmark for future, more advanced research on aspect-based sentiment analysis in academic review data.

XI. Conclusion

In this study, we developed and thoroughly evaluated a transformer-based baseline for sentence-level aspect-based sentiment analysis in academic peer reviews. By systematically fine-tuning ModernBERT and conducting a

targeted hyperparameter sweep, we achieved state-of-the-art performance on the ASAP-Review dataset, surpassing all classical and initial transformer baselines.

Interpretability analysis using LIME confirmed that the model's predictions are generally anchored in aspect-relevant and contextually meaningful words, supporting both transparency and reliability. Nonetheless, some misclassifications reveal the inherent challenges of modeling nuanced or ambiguous sentiment, particularly in complex academic texts.

While the project originally proposed the integration of advanced architectures such as target-aware encoding and CABiLSTM enhancements, the scope of this work focused on establishing and interpreting a strong transformer baseline. Future research should revisit these advanced strategies and explore larger datasets, alternative interpretability tools, and improved handling of subtle sentiment cues.

Overall, our results demonstrate that with careful optimization and analysis, a ModernBERT-based model can deliver robust and interpretable performance for ABSA tasks in academic peer review settings, providing a valuable benchmark for future advancements in this field.

REFERENCES

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect-Based Sentiment Analysis," in Proc. 8th International Workshop on Semantic Evaluation (SemEval), Dublin, Ireland, Aug. 2014, pp. 27–35.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[3] Z. Gao, J. Chen, X. Sun, and P. Zhou, "Target-Dependent Sentiment Classification With BERT," IEEE Access, vol. 7, pp. 154290–154299, 2019.

[4] B. He, R. Zhao, and D. Tang, "CABiLSTM-BERT: Aspect-Based Sentiment Analysis Model Based on Deep Implicit Feature Extraction," Knowledge-Based Systems, vol. 309, p. 112782, 2025.