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 BERT model. On top of this baseline, we will integrate two different enhancement techniques inspired by recent literature: (i) target-aware encoding strategies proposed for Target-Dependent Sentiment Classification with BERT, and (ii) deep implicit feature extraction methods from CABiLSTM-BERT. By adding these two mechanisms separately to the baseline, we will evaluate how each contributes to improving aspect-sentiment prediction performance. This experimental setup will allow us to systematically measure the effectiveness of both techniques and understand which factors contribute most to success in academic peer review sentiment analysis. In order to systematically progress the project, following research questions have been developed:

1) How can target-aware modifications to pretrained transformer models improve sentence-level aspect-based sentiment classification performance in academic peer reviews?

2) Can the integration of deep implicit feature extraction mechanisms further enhance the model's ability to associate sentiments with nuanced, contextually complex aspects compared to vanilla transformer baselines?

These questions evaluate both the effect of explicitly modeling aspect information and the impact of additional deep contextual modeling on ABSA outcomes.

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

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.

We will also combine the architectures to have both features from both papers. 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.

We anticipate that; Target-aware encoding will help the model focus attention on sentiment-bearing phrases linked to the correct aspect. Deep feature extraction will improve the model's robustness in handling long sentences, indirect sentiment, and complex sentence structures common in academic peer reviews. Both enhancements should provide measurable gains over the plain BERT fine-tuning baseline. When combined, they are expected to perform superior to either enhancement alone by simultaneously capturing both global contextual information and localized semantic cues around the aspects, leading to an overall improvement in aspect-sentiment prediction accuracy. Moreover, this work aims to demonstrate that different enhancement strategies can be successfully integrated into the same underlying architecture to achieve superior performance, emphasizing the flexibility and extensibility of transformer-based models for specialized tasks like sentence-level ABSA.

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.

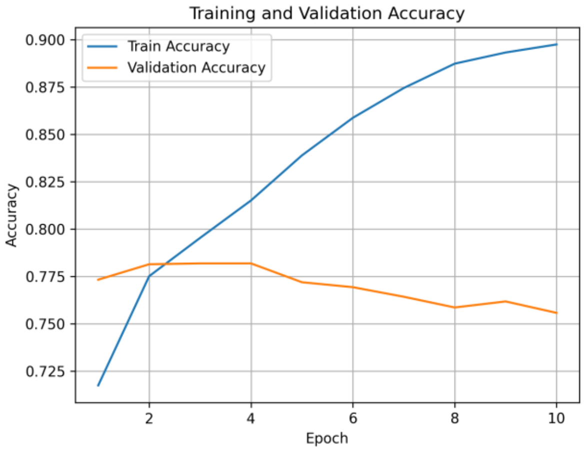
VI. Preliminary Results

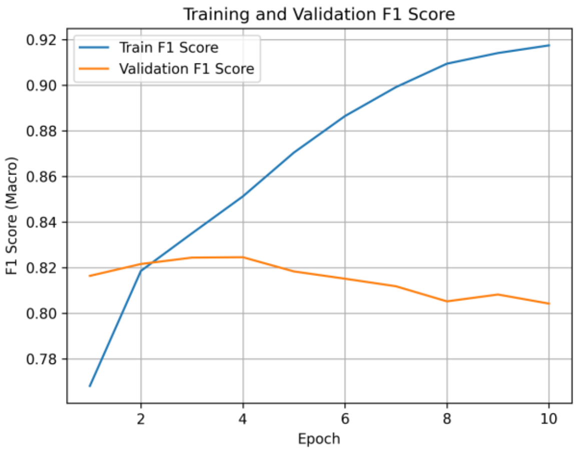
To evaluate our ModernBERT-based approach to aspect-based sentiment classification, we established and surpassed two predefined baselines: a majority class classifier and a TF-IDF + logistic regression model. We then trained and evaluated our initial deep learning model (ModernBERT) using the ASAP-Review 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.

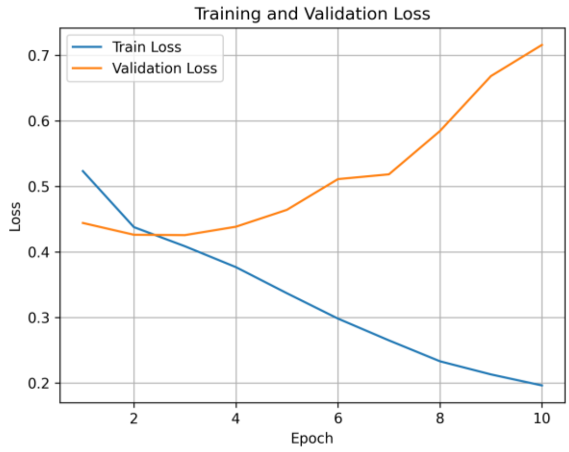
Before training, we processed the raw dataset from its original .jsonl format into a flat structured format compatible with model training. Each entry in the original file contains a sentence and a list of labeled aspect-sentiment spans. To standardize the format, we created a script that parses each line, extracts aspect information and corresponding sentiment, and writes them into a structured CSV file with the following columns: text, aspect, and sentiment.

This preprocessing step allows the training loop to easily associate aspect labels with corresponding sentences during tokenization and batching. It also ensures clean input handling for both baseline models and transformer-based fine-tuning. For other details please see the Github repo for training.

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.

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.