big data and
cognitive computing

Article

Explainable Aspect-Based Sentiment Analysis Using Transformer Models

Isidoros Perikos [1,2,3,*] and Athanasios Diamantopoulos [1]

[1] Computer Engineering and Informatics Department, University of Patras, 26504 Patras, Greece
[2] Computer Technology Institute and Press "Diophantus", 26504 Patras, Greece
[3] Electrical & Computer Engineering Department, University of Peloponnese, 26334 Patras, Greece
* Correspondence: perikos@ceid.upatras.gr

Abstract: An aspect-based sentiment analysis (ABSA) aims to perform a fine-grained analysis of text to identify sentiments and opinions associated with specific aspects. Recently, transformers and large language models have demonstrated exceptional performance in detecting aspects and determining their associated sentiments within text. However, understanding the decision-making processes of transformers remains a significant challenge, as they often operate as black-box models, making it difficult to interpret how they arrive at specific predictions. In this article, we examine the performance of various transformers on ABSA and we employ explainability techniques to illustrate their inner decision-making processes. Firstly, we fine-tune several pre-trained transformers, including BERT, RoBERTa, DistilBERT, and XLNet, on an extensive set of data composed of MAMS, SemEval, and Naver datasets. These datasets consist of over 16,100 complex sentences, each containing a couple of aspects and corresponding polarities. The models were fine-tuned using optimal hyperparameters and RoBERTa achieved the highest performance, reporting 89.16% accuracy on MAMS and SemEval and 97.62% on Naver. We implemented five explainability techniques, LIME, SHAP, attention weight visualization, integrated gradients, and Grad-CAM, to illustrate how transformers make predictions and highlight influential words. These techniques can reveal how models use specific words and contextual information to make sentiment predictions, which can improve performance, address biases, and enhance model efficiency and robustness. These also point out directions for further focus on the analysis of models' bias in combination with explainability methods, ensuring that explainability highlights potential biases in predictions.

Keywords: explainability; transformers; large language models; LIME; SHAP; attention visualization; integrated gradients; GRAD-CAM; aspect-based sentiment analysis

Citation: Perikos, I.; Diamantopoulos, A. Explainable Aspect-Based Sentiment Analysis Using Transformer Models. *Big Data Cogn. Comput.* **2024**, *8*, 141. https://doi.org/10.3390/bdcc8110141

Academic Editors: Zuchao Li, Min Peng and Yoichi Hayashi

Received: 31 July 2024
Revised: 28 September 2024
Accepted: 15 October 2024
Published: 24 October 2024

1. Introduction

A sentiment analysis is a natural language processing technique used to determine the attitude in textual data, often to understand customer opinions, feedback, or social media interactions. An aspect-based sentiment analysis (ABSA) is a specialized branch of a sentiment analysis that focuses on determining the sentiment expressed about specific aspects or attributes of a product, service, or entity within a given text. Unlike a traditional sentiment analysis, which aims to classify the overall sentiment of a text (e.g., positive, negative, or neutral), an aspect-based sentiment analysis provides a more granular analysis by identifying and evaluating sentiments associated with aspects [1]. For instance, in a product review, an aspect-based sentiment analysis can determine the sentiment expressed towards the product's features such as battery life, camera quality, or customer service, rather than just labeling the entire review as positive or negative.

An aspect-based sentiment analysis is a critical tool in natural language processing (NLP) that addresses the limitations of a traditional sentiment analysis by providing more granular insights. The main aim of an aspect-based sentiment analysis is to provide a detailed and nuanced understanding of opinions and sentiments expressed in textual data. This is particularly valuable in various domains such as product reviews, social media monitoring, customer feedback analyses, and market research [2]. While a general sentiment analysis focuses on identifying the overall sentiment of a piece of text, ABSA delves deeper by associating sentiments with specific aspects or attributes of entities mentioned in the text. This fine-grained approach is essential in real-world applications where understanding the sentiment towards individual product features, services, or components is necessary. So, the need for an aspect-based sentiment analysis (ABSA) arises from the limitations of a traditional sentiment analysis, which often provides only an overall sentiment (positive, negative, or neutral) for a given text. However, in many real-world scenarios, sentiments are expressed towards specific aspects or features rather than the entire entity. By identifying sentiments at the aspect level, businesses and organizations can gain actionable insights into specific strengths and weaknesses of their products or services. For example, a company can identify that while their product is generally well received, customers are dissatisfied with the battery life, allowing them to make targeted improvements. For example, in customer reviews of a product, a user might express satisfaction with the battery life but dissatisfaction with the design. Without ABSA, such detailed insights are lost, and companies are left with an incomplete understanding of customer feedback. ABSA addresses this issue by identifying and analyzing sentiments at a more granular level, allowing for a precise understanding of how users feel about specific aspects of products, services, or experiences. This capability is crucial in industries like e-commerce, hospitality, and social media analyses, where organizations must focus on both the overall sentiment and the sentiments associated with individual aspects. By improving the ability to capture feedback, ABSA enables businesses to make more informed decisions, improve customer satisfaction, and target specific areas for improvement. For instance, in customer reviews of electronic products, ABSA allows companies to extract sentiments tied to features like battery life, camera quality, or customer support, offering actionable insights for targeted improvements. Similarly, in the hospitality industry, ABSA enables hotels and restaurants to assess feedback on individual aspects such as service quality, ambiance, or pricing, thereby providing a clearer understanding of areas requiring attention. In financial services, ABSA can analyze sentiment towards specific elements of a service, such as interest rates, mobile app features, or customer service interactions, helping institutions better respond to customer needs. By offering these detailed insights, ABSA plays a pivotal role in decision-making processes across various industries, highlighting its necessity for businesses seeking to remain competitive in a data-driven world.

Accurate and efficient aspect-based sentiment analysis methods are quite important and desirable for many reasons. First, they enhance the accuracy and relevance of sentiment analyses by considering the context in which sentiments are expressed. This context awareness is crucial in understanding the true sentiment of a text, as the same word or phrase can convey different sentiments depending on the aspect being discussed. Second, they can provide more detailed insights, enabling organizations to make data-driven decisions and address specific issues that matter to their customers. Finally, with the proliferation of user-generated content on the internet, they can effectively process and analyze large volumes of text data, making it a vital tool for businesses seeking to maintain a competitive edge in today's data-driven world [3]. So, an aspect-based sentiment analysis has become crucial for effectively interpreting and understanding opinions and public stances towards events, products, and services and accurate methods are highly desirable [4].

Deep learning techniques and transformers have revolutionized the field of sentiment analyses [5]. Deep learning methods have greatly improved the performance and the capabilities of systems by automatically learning features from textual data [6]. Indeed, one of the most significant breakthroughs has been the development of transformers. Introduced by Vaswani [7], transformers have become the backbone of many state-of-the-art natural language processing (NLP) models [8]. Transformers leverage self-attention mechanisms to process and understand the relationships between words in a sentence, regardless of their position. This allows transformers to capture long-range dependencies and context more effectively than previous models, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Transformers have paved the way for several powerful pre-trained language models, such as BERT (Bidirectional Encoder Representation from Transformers), DistilBERT, ALBERT, RoBERTa, and XLNet. These models are pre-trained on large corpora of text data to learn general language representations, which can then be fine-tuned for specific tasks like ABSA. Fine-tuning involves training the pre-trained model on a smaller, task-specific dataset, allowing it to adapt its knowledge to the nuances of the target task. The ability of transformers to leverage large-scale pre-training and fine-tuning has led to significant improvements in a wide range of tasks, including an aspect-based sentiment analysis.

In this article, we examine the performance of various transformers on ABSA and we employ explainability techniques to illustrate their inner decision-making processes. Firstly, we fine-tune several pre-trained transformers, including BERT, ALBERT, RoBERTa, DistilBERT, and XLNet, on extensive datasets such as MAMS, SemEval, and Naver. These datasets consist of over 16,100 complex sentences, each containing at least two aspects and corresponding polarities. The models were fine-tuned using optimal hyperparameters, and RoBERTa achieved the highest accuracy of 89.16% on SemEval and MAMS and 97.62% on Naver. After that, to shed light on the decision-making processes of trans-formers, we implemented five explainability techniques: LIME, SHAP, attention weight visualization, integrated gradients, and Grad-CAM. These techniques provide valuable insights into how transformers make predictions and highlight influential words and phrases. Also, they illustrate the inner workings of transformer models, showing how they utilize specific words and contextual information to make sentiment predictions.

The article is structured as follows. Section 2 provides a comprehensive review of recent works on aspect-based sentiment analyses, highlighting the challenges and the limitations of existing methods. Also, it provides the context and the motivation for our study. Section 3 outlines our methodology, detailing the datasets used, which span different domains. It also elaborates on the fine-tuning process of transformer models and explains how explainability is applied to interpret the model outputs at the aspect level. Section 4 presents an in-depth explanation of the explainability techniques utilized in the study, discussing how they generate explanations for predictions and the insights they offer into model behavior. Section 5 discusses the main results and findings of the study, offering insights into how explainability techniques operate and can provide transparency in real-world applications. Finally, Section 6 concludes the article and outlines potential future research directions.

2. Related Works

In the work presented in [9], the authors introduced T-MGAN, a model for an aspect-based sentiment analysis with four layers: Input Word Embedding, Intra-Feature Extraction, Inter-Feature Multi-Attention, and Output. The model takes a sentence and its aspects as input and classifies sentiment polarity (positive, negative, or neutral) for each aspect. GloVe embeddings are used in the Input Layer. The Intra-Feature Layer uses transformer and tree transformer encoders for word-level and phrase-level representations. The Inter-Feature Layer employs a multi-attention mechanism to capture context–aspect interactions. The Output Layer concatenates these representations and uses a SoftMax layer for sentiment classification, trained with cross-entropy loss and L2 regularization. T-MGAN achieved 76.38% accuracy and 73.02% Macro-F1 on the laptop dataset, 82.06% accuracy and 72.65% Macro-F1 on the restaurant dataset, and 71.23% accuracy and 70.63% Macro-F1 on the Twitter dataset.

In the work in [10], the authors proposed the FGAOM model for Aspect–Opinion Mining, comprising five main components: BERT domain adaptation, context feature extraction, long-term dependency learning, aspect feature capturing, and polarity classification. The model fine-tunes BERT with domain-specific corpora to create DA-BERT, capturing local and global context embeddings. These embeddings are processed through a dynamic masking layer, a fusion layer, and a fine-tuning layer with Bidirectional GRU, Multi-Head Self-Attention, and convolutional layers. Finally, the model uses a fully connected layer with an MLP and SoftMax function for the classification. FGAOM achieves 85.24% accuracy and an 83.67% F1 score on the laptop dataset, 91.6% accuracy and an 88.01% F1 score on the restaurant dataset, and 80.6% accuracy and a 79.21% F1 score on the Twitter dataset.

In [11], the authors proposed the SA-EXAL model, which enhances a pre-trained BERT-base model with syntactically aware self-attention mechanisms. This model integrates dependency relations into the self-attention mechanism, adding syntactically aware heads in parallel to BERT during fine-tuning and testing stages. This allows leveraging both external syntactic information and BERT's pre-trained knowledge. The SA-EXAL model improves sentiment prediction across domains, achieving F1 scores of 47.59 for the aspect sentiment (AS) and 75.79 for the overall sentiment (OP) in the target domain "laptop" when trained on "restaurant." It also achieved 40.50 (AS) for "Device", 54.67 (AS) and 80.85 (OP) for "restaurant", and 42.19 (AS) for "Device" and 54.54 (AS) for "restaurant" when trained on "Device".

Authors in [12] proposed the HSAN model with six main components: the Word Embedding Layer, Word and Sentence Encoder Layer, Global Context-Aware Word Representation (GCWR) Layer, Information Fusion Layer, Word-Specific Context-Aware Representation (WSCR) Layer, and Output Layer. The model uses bidirectional LSTM for contextual word representations and attention mechanisms for assigning importance weights. A self-attention mechanism identifies internal dependencies among words, and a Conditional Random Field (CRF) predicts aspect-term tags. HSAN achieved precision = 88.43, recall = 80.04, and F-Score = 84.01 on the laptop dataset, and precision = 90.24, recall = 89.70, and F-Score = 89.96 on the restaurant dataset.

In the work in [13], the authors proposed the MDAE-BERT model, which fine-tunes a pre-trained BERT model for aspect extraction using labeled data from multiple domains. It employs an IOB (Inside, Outside, Beginning) token classification framework. The source domain (Ds) has labeled reviews, while the target domain (Dt) has unlabeled reviews. BERT encodes review token sequences to generate contextual word embeddings and predict masked tokens. For aspect extraction, the model is fine-tuned with token classification, classifying tokens into IOB tags through a dense layer and SoftMax, optimized with cross-entropy loss. MDAE-BERT showed a minimum performance increase of 7.99% for F1-Macro and 10.62% for accuracy over LSTM. It achieved 65.86% F1-Macro for the laptop domain, nearly 90% better than LSTM.

In [14], the authors proposed the DualGCN model, which integrates syntactic and semantic information using either BiLSTM or BERT for encoding. It has two main modules: SynGCN for syntactic dependencies and SemGCN for semantic relationships via self-attention. A BiAffine module facilitates information exchange between these modules. The final aspect representation is used for sentiment prediction. The model employs orthogonal and differential regularizers and minimizes cross-entropy loss. Experimental results show that DualGCN achieves 78.48% accuracy and 74.74% Macro-F1 on the laptop dataset, 84.27% accuracy and 78.08% Macro-F1 on the restaurant dataset, and 75.92% accuracy and 74.29% Macro-F1 on the Twitter dataset.

Authors in [15] proposed the dotGCN model that uses a recognition network to create an opinion tree from an input sentence and aspect term, capturing structural relationships. Multi-layered Graph Convolutional Networks (GCNs) are applied to BERT output vectors to model these relations and extract aspect-specific features. An attention-based model then learns sentiment polarity. The model is trained using reinforcement learning for opinion tree induction and backpropagation for classification, incorporating various loss functions. Experimental results show that dotGCN achieves 84.95% accuracy and 84.44% F1 on the MAMS dataset, 78.11% accuracy and 77.00% F1 on SemEval Twitter, 81.03% accuracy and 78.10% F1 on the laptop dataset, 86.16% accuracy and 80.49% F1 on Rest14, 85.24% accuracy and 72.74% F1 on Rest15, and 93.18% accuracy and 82.32% F1 on Rest16.

Authors in [16] proposed the KDGCN model, which includes a sentence encoder, a knowledge enhancement module, a semantic learning module, a syntax-aware module, and a sentiment classifier. It uses embeddings from Glove, BERT, or BiLSTM, enhanced with SenticNet sentiment vectors and aspect-related words, and employs Graph Convolutional Networks (GCNs) for semantic and syntactic information. The model combines these for sentiment polarity. It achieves 79.00% accuracy and 75.03% macro-F1 on laptop14, 84.91% accuracy and 78.48% macro-F1 on restaurant14, 82.10% accuracy and 67.13% macro-F1 on restaurant15, and 90.74% accuracy and 73.46% macro-F1 on restaurant16.

In [17], the authors proposed the BERT-XGBoost model for an aspect-based sentiment analysis on customer reviews. The model preprocesses data, uses BERT to encode semantic information, and employs transformer encoders and an attention mechanism to capture relationships between context and aspect. XGBoost then predicts sentiment polarity. Results show that bert-xgboost achieves 85.01% accuracy and 78.9% macro-F1 on laptop14, 87.86% accuracy and 81.64% macro-F1 on restaurant14, 86.7% accuracy and 79.77% macro-F1 on restaurant15, and 93.71% accuracy and 80.37% macro-F1 on restaurant16.

Authors in [18] proposed the LDEGCN model for an aspect-based sentiment analysis. The model uses a multi-layer attention structure and graph neural networks to capture syntactic and semantic information, integrating external sentiment knowledge and dependency types. It includes an embedding layer, a semantic feature extraction, a feature fusion, and a sentiment classifier. BERT initializes aspect-aware word vectors, and a multi-layer attention mechanism to extract semantic features. The model fuses information using a Graph Convolutional Network for sentiment classification. Results show 81.25% accuracy and 78.17% macro-F1 on laptop14, 86.34% accuracy and 81.16% macro-F1 on restaurant14, 85.42% accuracy and 72.05% F1 on restaurant15, 91.56% accuracy and 79.45% macro-F1 on restaurant16, and 76.43% accuracy and 75.22% F1 on Twitter.

Authors in [19] proposed the SDTGCN method for a sentiment analysis, leveraging sentence information like aspect positions, sentiment relationships, POS tags, and dependency distances. It constructs two dependency weight matrices: the adjacency-enhanced and the sub-adjacent dependency weight matrices. These matrices emphasize relevant nodes and capture sentiment propagation. The weighted GCN aggregates node information, assigning higher weights to important nodes, while Bi-LSTM embeddings capture contextual information. An attention mechanism refines the analysis by focusing on aspect–context relationships. The model is optimized using cross-entropy loss and gradient descent. Experimental results show that SDTGCN achieves 78.64% accuracy and 75.50% F1 on laptop14, 83.82% accuracy and 76.13% F1 on restaurant14, 83.21% accuracy and 67.07% F1 on restaurant15, 91.53% accuracy and 77.08% F1 on restaurant16, and 76.25% accuracy and 74.59% F1 on the Twitter dataset.

Authors in [20] proposed the IDGNN model, which uses BERT for word embeddings, followed by BiLSTM for contextual representation, and employs RGAT for syntactic dependencies and GCN with attention for semantic associations. These are combined through a gated fusion mechanism and a fully connected layer for sentiment classification. Results show that IDGNN achieves 78.07% accuracy and 74.17% macro-F1 on the laptop, 83.40% accuracy and 76.37% macro-F1 on the restaurant, 81.80% accuracy and 80.90% macro-F1 on MAMS, and 75.94% accuracy and 74.37% macro-F1 on Twitter, with BERT embeddings; IDGNN scores higher: 81.12% accuracy and 77.73% macro-F1 on the laptop, 87.25% accuracy and 81.16% macro-F1 on the restaurant, 84.57% accuracy and 83.42% macro-F1 on MAMS, and 76.72% accuracy and 75.92% macro-F1 on Twitter.

In recent years, large language models like GPT have shown significant potential in ABSA. The authors of the paper in [7] explore the application of GPT-3.5 in ABSA, particularly focusing on its few-shot learning capabilities. The proposed method employs prompt engineering to guide GPT in performing core ABSA tasks, such as aspect extraction, sentiment classification, and opinion–target pairing, with minimal labeled data. By designing specific prompts, the model can generate predictions with little to no task-specific training, making it highly adaptable to real-world situations where annotated datasets are scarce. The study evaluates the impact of manual prompts and auto-generated prompts to optimize GPT's performance in few-shot settings. Despite the flexibility and generalization ability of GPT, the study finds that fine-tuned models like BERT consistently outperform GPT in structured ABSA tasks that demand deep domain understanding, such as precise aspect–opinion extraction. While GPT demonstrates strong generalization, it struggles with tasks requiring more nuanced context comprehension. The results suggest that while GPT's few-shot capabilities are useful, especially in scenarios with limited labeled data, models fine-tuned for specific tasks, like BERT, provide better accuracy and consistency.

Another significant advancement in the application of transformer models to ABSA is presented in [21]. The authors propose a novel approach that integrates BERT, post-trained specifically for the ABSA task, with an Interactive Attention Network (IAN) to enhance the model's ability to capture aspect–target interactions and sentiment dependencies. In this method, BERT is fine-tuned on domain-specific data to improve its understanding of the aspect-level sentiment, while the IAN component is designed to refine the interaction between the aspect and its corresponding context in the text. The attention mechanism used in IAN allows the model to focus on the most relevant parts of the sentence for both the aspect and the sentiment, leading to more precise sentiment predictions. This hybrid architecture aims to leverage the strengths of BERT's pre-training with the fine-grained attention mechanism of IAN to improve ABSA performance. The paper demonstrates that IAN-BERT outperforms standard BERT models in both sentiment classification and aspect extraction, showing the importance of combining attention mechanisms with transformer-based models for better handling of aspect-level tasks.

In [22], the authors explore the effectiveness of ChatGPT in performing ABSA through prompt-based techniques. The paper evaluates ChatGPT's ability to extract aspects and their corresponding sentiment polarities using zero-shot and few-shot learning. To assess ChatGPT's capabilities, the authors design specific prompts aimed at instructing the model to identify sentiment and opinion targets without requiring extensive task-specific training. This approach allows ChatGPT to generalize across different datasets with minimal labeled data. The study compares ChatGPT's performance to fine-tuned models like BERT, highlighting ChatGPT's strength in general sentiment understanding. However, the authors find that BERT and other fine-tuned models still outperform ChatGPT in structured tasks that require domain-specific knowledge, particularly in cases where aspect–opinion pairs are crucial. While ChatGPT performs well in zero-shot settings, its performance in ABSA tasks often lacks the precision and depth achieved by fine-tuned models, emphasizing the importance of task-specific training for more nuanced sentiment analyses.

The authors in [23] present a comprehensive evaluation of GPT models, particularly GPT-3.5, for a sentiment analysis. The study proposes three key strategies to enhance a sentiment analysis: prompt engineering, fine-tuning, and embedding classification. First, the authors employ prompt engineering to guide GPT-3.5 in performing a sentiment analysis by crafting task-specific prompts that extract sentiment information directly from the text. This method leverages GPT-3.5′s ability to generalize from minimal task-specific training, making it efficient in handling sentiment tasks with limited labeled data. Second, the study fine-tunes GPT models on labeled datasets to enhance performance, allowing the model to adapt its pre-trained knowledge to specific sentiment analysis tasks. Finally, an innovative approach to embedding classification is introduced, focusing on optimizing sentiment representation within the GPT framework. The results show a significant improvement over state-of-the-art sentiment analysis methods, with an increase of more than 22% in the F1 score compared to traditional models. Moreover, the paper addresses common challenges in sentiment analyses, such as handling complex language constructs, context, and sarcasm, demonstrating the enhanced capabilities of GPT models in these areas.

Despite the significant advancements in an aspect-based sentiment analysis (ABSA), existing methods still face several limitations that justify the need for further research [24]. Traditional ABSA models often struggle with handling complex sentences containing multiple aspects with varying sentiments. This challenge is particularly pronounced in real-world applications where nuanced sentiment detection is crucial. Moreover, many current models rely heavily on pre-defined aspect categories, limiting their flexibility and adaptability to new domains or datasets. Another critical limitation is the lack of explainability in many state-of-the-art models. While models like BERT and RoBERTa have demonstrated impressive performance in ABSA tasks, their decision-making processes remain opaque, making it difficult for users to trust and interpret their predictions. Additionally, existing methods may suffer from biases in data, which can lead to skewed sentiment predictions, especially in cases involving minority groups or less frequent aspects. Our research addresses these gaps by integrating advanced transformer models with explainability techniques, aiming to enhance both the accuracy and interpretability of ABSA models. By providing clearer insights into how models arrive at their predictions, we hope to improve user trust and make the models more applicable in real-world scenarios where transparency and adaptability are key.

3. Transformer Model Fine-Tuning

In this section, we outline the process we followed. First, we present the dataset collection used in the study and analyze the characteristics of the data. Then, we present the fine-tuning process of the pre-trained transformer models for the aspect-based sentiment analysis task, and we report their performance. After that, we present the implementation and use of explainability techniques and the way that each technique shed light on the decision-making procedure of the fine-tuned transformer models and how they utilize words and contextual content in sentences. Figure 1 illustrates the overall workflow of our aspect-based sentiment analysis (ABSA) framework.

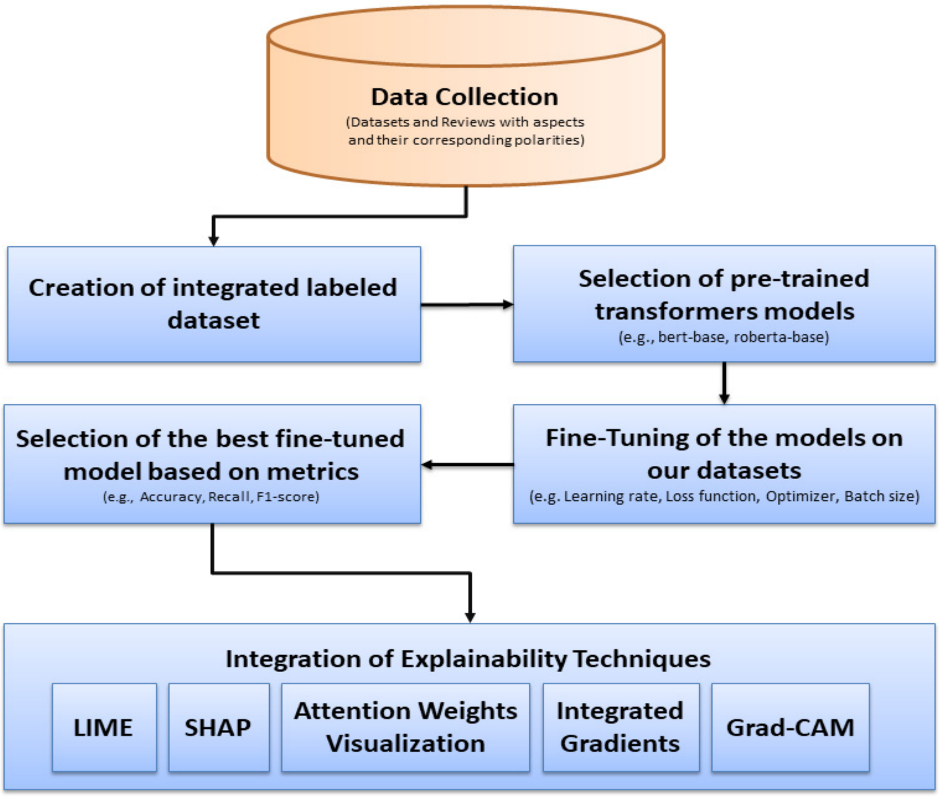

Data Collection
(Datasets and Reviews with aspects and their corresponding polarities)
Creation of integrated labeled dataset
Selection of pre-trained transformers models
(e.g., bert-base, roberta-base)
Fine-Tuning of the models on our datasets
(e.g. Learning rate, Loss function, Optimizer, Batch size)
Selection of the best fine-tuned model based on metrics
(e.g., Accuracy, Recall, F1-score)
Integration of Explainability Techniques
LIME
SHAP
Attention Weights Visualization
Integrated Gradients
Grad-CAM

Figure 1. The overall workflow of our framework.

The process begins with data collection, where reviews containing labeled aspects and their corresponding sentiment polarities are gathered. We then create a labeled dataset and fine-tune pre-trained transformer models (such as BERT and RoBERTa) on this data. After fine-tuning, we evaluate the models using performance metrics (e.g., accuracy, F1 score) and select the best-performing model. Finally, explainability techniques (LIME, SHAP, attention weight visualization, integrated gradients, Grad-CAM) are applied to the selected model to provide insights into how the model makes predictions at the aspect level. This diagram provides a visual guide to the methodological steps outlined in this section, from data preparation to explainability analyses.

3.1. Study Design and Datasets Utilized

In the context of the study, and for fine-tuning the pre-trained transformer models, we utilized labeled datasets in an aspect-based sentiment analysis, having their aspects and the respective aspect polarities defined. Specifically, five widely used and diverse datasets were selected and used, which are the MAMS dataset [25], the SemEval 2014 Task 4: laptop reviews and restaurant reviews dataset [26], the SemEval 2015 Task 12: restaurant reviews dataset [27], the SemEval-2016 Task 5: restaurant reviews dataset [28], the ACOS dataset [29], and the Naver Labs dataset.

The MAMS dataset has more than 4297 sentences and 11,186 aspects. It is a large-scale multi-aspect dataset. MAMS is a challenging dataset for an aspect-based sentiment analysis, in which each sentence contains at least two aspects with different sentiment polarities.

The SemEval 2014 Task 4 dataset is designed for an aspect-based sentiment analysis and includes laptop and restaurant reviews. The laptop reviews dataset contains 3845 sentences, while the restaurant reviews dataset contains 3.041 sentences. These reviews focus on various aspects of laptops and restaurants. Each review is annotated with specific aspect terms and the corresponding sentiment polarity (positive, negative, neutral). In the laptop reviews, consumers discuss features such as battery life, screen quality, performance, design, and price. For example, they might mention the long battery life of a particular model or criticize the screen resolution. Each aspect mentioned in the reviews is tagged with the sentiment expressed towards it.

The SemEval 2015 Task 12 dataset also focuses on an aspect-based sentiment analysis specifically for restaurant reviews. It contains 2541 sentences from customer reviews of various restaurants. Each review includes text discussing different aspects of the dining experience, such as food quality, service, ambiance, price, and location. The dataset is annotated with aspect terms and their corresponding sentiment polarity.

The SemEval-2016 Task 5 dataset is also focused on an aspect-based sentiment analysis for restaurant reviews. This dataset includes 3500 sentences from customer reviews of various restaurants. Each review contains text that discusses different aspects of the dining experience, such as food quality, service, ambiance, price, and location. The dataset is annotated with aspect terms and their corresponding sentiment polarity (positive, negative, neutral). In these restaurant reviews, customers provide detailed feedback on specific aspects. For instance, a review might praise the excellent taste of the food, criticize the unfriendly service, or comment on the restaurant's pleasant ambiance. Each mentioned aspect is tagged with the sentiment expressed towards it.

The ACOS dataset consists of the restaurant and the laptop parts and extends the existing SemEval restaurant dataset by adding annotations for implicit aspects, implicit opinions, and quadruples. Also, the laptop part is collected from the Amazon laptop domain and includes annotations for quadruples encompassing all explicit and implicit aspects and opinions.

The Naver Labs dataset consists of user reviews from Foursquare, manually annotated according to the SemEval2016 guidelines for the restaurant domain. From the approximately 215K English-language reviews, a sample of 585 reviews containing 1006 sentences was randomly selected for annotation. This dataset provides a real-world benchmark to evaluate the models' robustness beyond our dataset.

These datasets were combined, and a larger dataset was created containing more than 16,100 text sentences, where each sentence has at least one aspect along with the corresponding polarity. More specifically, we had 30,813 aspects in total, of which there are 14,455 positive polarities, 9261 negative polarities, and 7097 neutral polarities. These data were used to fine-tune pre-trained transformer models used and evaluate their performance. In the following subsections, the experimental setups and the collected results are thoroughly presented.

3.2. Transformer Models and Fine-Tuning Process

We investigated the eight pre-trained transformer models that are the BERT, the DistilBERT, the ALBERT, the RoBERTa, the XLNet, and their variations. Specifically, we utilized BERT-base, BERT-large [30], DistilBERT-base [31], ALBERT-base [32], RoBERTa-base, RoBERTa-large [33], and XLNet-base [34].

BERT-base-uncased is a transformer model designed to understand the context of words in a sentence by looking at both directions (left-to-right and right-to-left) simultaneously to understand the full context of a word. It has 12 layers, each with a hidden size of 768 and 12 attention heads, totaling 110 million parameters. It is trained on pairs of sentences to understand the relationship between them.

BERT-large-uncased expands upon BERT-base with 24 layers, each having a hidden size of 1024 and 16 attention heads, totaling 340 million parameters. It has higher capacity and complexity due to its larger number of layers, hidden units, and parameters and it is designed to capture more detailed and complex patterns in the data, leading to potentially better performance on complex tasks.

DistilBERT-base-uncased is a smaller, distilled version of BERT. It is created via knowledge distillation, where a smaller model is trained to reproduce the behavior of a larger model. DistilBERT has 6 layers, a hidden size of 768, and 12 attention heads, resulting in 66 million parameters in total.

ALBERT-base-v1 is designed to reduce the model size of BERT and increase training efficiency. It employs parameter sharing and factorized embedding parameterization to achieve a compact model with 12 layers, each having a hidden size of 768 and 12 attention heads, summing to 12 million parameters. ALBERT uses parameter-reduction techniques that allow for large-scale configurations, overcome previous memory limitations, and report good behavior with respect to model degradation.

ALBERT-large-v1 is a larger model of ALBERT providing more capacity for complex tasks. It has 24 layers, each with a hidden size of 1024 and 16 attention heads. It employs parameter efficiency across layers and has 18 million parameters in total. It shares parameters across layers, assisting in avoiding overfitting of the model.

RoBERTa-base is a robustly optimized BERT having an optimized pre-training process including training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data. It has 12 layers, a hidden size of 768, and 12 attention heads and a total of 125 million parameters. Unlike the static masking of the BERT model, it uses dynamic masking, applying different masks to each input sequence.

RoBERTa-large is an enhanced version of RoBERTa-base, featuring 24 layers, a hidden size of 1024, and 16 attention heads, summing to 355 million parameters. RoBERTa-large benefits from the same optimizations in the pre-training process; however, it has the potential to better leverage these improvements due to its larger architecture.

XLNet-base-cased is an autoregressive pre-training model with 12 layers, each with a hidden size of 768 and 12 attention heads, totaling 110 million parameters. It is designed to overcome limitations of BERT by integrating the strengths of autoregressive (AR) and autoencoding (AE) models and leverages a Transformer-XL as its backbone for better handling of long-term dependencies.

GPT-3.5-turbo is a variant of OpenAI's GPT-3.5 model, optimized for faster performance and lower cost. It is based on the same transformer architecture as previous GPT models, such as GPT-3, which had 175 billion parameters. GPT-3.5-turbo is likely smaller than GPT-4 but still highly capable of handling a wide range of natural language tasks. Unlike models like BERT, which are bidirectional and suited for tasks like classification, GPT-3.5-turbo is an autoregressive model, generating text in one token at a time.

GPT-4 is a large language model built on the transformer architecture, designed to process and generate human-like text by understanding the context within a given input. Unlike BERT, which primarily focuses on understanding text bidirectionally for tasks like classification and sequence labeling, GPT-4 is autoregressive, excelling in text generation, language understanding, and task adaptation. It has a vast number of parameters and has been trained on a diverse and extensive corpus, allowing it to handle a wide variety of tasks without task-specific fine-tuning.

3.3. Transformer Fine-Tuning

In the fine-tuning process, we first split the dataset into training, validation, and test sets, with respective proportions of 70%, 15%, and 15%. To ensure that each model receives the input in a format that maximizes its ability to learn the relationships between the text and its corresponding aspects, we tokenize the text in a specific manner. For each sentence, we concatenate the text with its corresponding aspect using the appropriate separator token designated by the pre-trained models (e.g., [SEP] for BERT, <sep> for XLNet). This token effectively demarcates the main text from the aspect, allowing the model to distinctly understand the relationship between the two parts during training. The use of these separator tokens is crucial because it enables the transformer models to handle the input as a pair of sequences rather than a single concatenated string. This separation ensures that the models can focus on the sentiment expressed specifically in relation to the aspect in question, rather than making a generalized prediction based on the entire text. By maintaining the integrity of the sequence-pair structure, we allow the models to leverage their pre-trained capabilities more effectively, improving the precision of aspect-based sentiment predictions.

We conducted several experiments to train our models, varying the learning rate and batch size while maintaining a constant number of epochs at 15, without altering the loss function or the optimizer. Through these experiments, we identified optimal values for each hyperparameter. Specifically, a grid search was performed to identify the best combination of these parameters of our models.

Regarding the learning rate, we experimented with a wide range of different learning rates and found that a learning rate of 2×10^{-5} provided the best balance between convergence speed and model performance. Lower learning rates allowed for more precise optimization but required more epochs, while higher learning rates led to unstable training.

Regarding the batch size, we experimented with a wide range of sizes in the study. In general, larger batch sizes can lead to faster training while smaller batch sizes allow for more frequent updates to the model but can slow down training. Our experiments showed that a batch size of 16 strikes the best balance between training speed and memory usage for base transformer models, while for large versions, a larger batch size up to 128 assisted greatly in their efficient fine-tuning process.

An epoch refers to one complete pass through the entire training dataset. We fixed the number of epochs at 15 after determining that this value was sufficient for the model to learn effectively from the data. Also, we noted that training for more epochs did not significantly improve performance and sometimes tended to lead to overfitting issues.

For the loss function, we used the cross-entropy sequence classification tasks. The cross-entropy loss increases as the predicted probability diverges from the actual label, thus penalizing incorrect predictions more heavily and driving the model to produce accurate classifications by minimizing this loss during training. This characteristic makes cross-entropy loss particularly effective and popular in training deep learning models for various classification tasks.

Regarding the optimizer, we experimented with various optimizers and the optimizer that assisted the models to report their best performance was the AdamW optimizer. It is an extension of the Adam optimizer with weight decay, which helps prevent overfitting by penalizing large weights. It is efficient and effective in training deep neural networks and combines the benefits of adaptive learning rates with weight decay regularization.

The AdamW optimizer helped our fine-tuned transformer models to face overfitting by penalizing large weights, leading to better generalization on unseen data. The optimizer retains the benefits of Adam, such as adaptive learning rates and momentum, and is a quite effective choice for fine-tuning pre-trained transformer models in natural language processing tasks.

In summary, through our experiments, we specified proper values for the hyperparameters as follows: using dynamic learning rates and specifying the best performance of the models to be a learning rate of 2×10^{-5} and a batch size in the range 8–128, and reporting the best batch size of each transformer model, 15 epochs, using cross-entropy loss and the AdamW method as an optimizer. These settings yielded the best performance and efficiency for our fine-tuning process for the aspect-based sentiment analysis task and below we present the performance results of the fine-tuned transformer models.

Regarding the GPT-4 and GPT 3.5-turbo, in our experiments, we used them through the OpenAI API. We utilized GPT-4′s and GPT-3.5-turbo's general pre-trained capabilities to predict the sentiment of aspects within the given text. By sending prompts to GPT-4 and GPT-3.5-turbo via the API, we evaluated their performance on various datasets, including the Naver and MAMS Semeval datasets.

3.4. Performance Results

3.4.1. Results on MAMS and SEMEVAL Datasets

In the following, we present the main results obtained from the various experiments performed on the pre-trained models on the MAMS and the SemEval. Also, for further explanation, in the training set, there are 21,599 aspects of which 10,111 are positive, 6503 are negative, and 4985 are neutral. In the validation set, there are 4610 aspects of which 2145 are positive, 1373 are negative, and 1092 are neutral. Finally, in the test set, there are 4604 aspects of which 2199 are positive, 1385 are negative, and 1020 are neutral. More specifically, for each base pre-trained model, we experimented with a variety of learning rate and batch size options while keeping the epoch fixed at 15 based on the results we collected during the training phase. Thus, for each such model that we fine-tuned with various choices in hyperparameters, we chose from each class of pre-trained models the one that had the best metrics: the accuracy, precision, recall, and F1 score. Specifically, to evaluate the performance of the models used in our study, we employed several well-established metrics that are commonly used in sentiment analysis and machine learning tasks. Specifically, we measured accuracy, precision, recall, and the F1 score. Accuracy provides an overall measure of how often the model correctly predicts the sentiment for a given aspect, while precision indicates the proportion of positive predictions that were correct. Recall measures the model's ability to correctly identify all relevant positive cases, and the F1 score provides a harmonic mean of precision and recall, offering a balanced view of performance, especially in cases with imbalanced data. In addition, we conducted cross-validation to ensure the robustness of our results and reported the performance on test data to reflect the model's generalizability. These metrics are detailed in the manuscript, along with comparative performance tables for each of the transformer models we tested, enabling readers to clearly understand and assess the effectiveness of each model. The performance results of the transformer models are presented in Table 1.

Based on the fine-tuning results, roberta-base achieved the highest performance among all models with an accuracy of 89.16%, precision of 87.94, recall of 88.23, and an F1 score of 88.08. In contrast, distilbert-base-uncased had the lowest accuracy at 85.95%. When comparing roberta-base to roberta-large where they fine-tuned with exactly the same hyperparameters, the base model outperformed the large model across all metrics, suggesting that the base model was more effective for this task despite the large model's increased capacity. For the ALBERT models where they also fine-tuned with the same hyperparameters, albert-large-v1 outperformed albert-base-v1 in the accuracy (86.38% vs. 85.49%), precision (85.73 vs. 85.10), recall (84.31 vs. 83.31), and F1 score (84.80 vs. 84.00), indicating that the increased capacity of the large model provided better performance. Overall, the results highlight the significance of both model architecture and size, with roberta-base emerging as the top performer, and the importance of appropriately tuning hyperparameters and considering model complexity relative to the specific task and dataset.

Table 1. The performance results of the best fine-tuned models on MAMS and SEMEVAL datasets. The best model is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 Score
bert-base-uncased	86.64%	85.64	85.06	85.30
bert-large-uncased	87.14%	86.31	84.62	85.32
distilbert-base-uncased	85.95%	84.94	84.21	84.55
albert-base-v1	85.49%	85.10	83.31	84.00
albert-large-v1	86.38%	85.73	84.31	84.80
roberta-base	**89.16%**	**87.94**	**88.23**	**88.08**
roberta-large	88.86%	87.65	87.63	87.63
xlnet-base-cased	88.68%	87.31	87.72	87.51
Bi-LSTM	73.18%	71.65	69.83	70.55
LSTM	71.92%	70.04	70.81	70.37
RNN	60.38%	58.54	58.69	57.92
GPT-4	78.70%	74.73	73.47	73.52
GPT-3.5-turbo	73.89%	67.45	64.69	61.39

In addition to the transformer models, we implemented traditional deep learning models such as Long Short-Term Memory (LSTM), bidirectional Long Short-Term Memory (bi-LSTM), and RNN models. The Bi-LSTM is a type of Recurrent Neural Network (RNN) that processes input sequences in both forward and backward directions, allowing it to capture context from both preceding and following words in a sentence. This makes it particularly effective for tasks like a sentiment analysis, where understanding the full context is crucial. For our baseline, we used the same dataset and preprocessing steps as those used for the transformer models. In addition, the Bi-LSTM model was trained using a custom tokenization process where both the input sentence and the aspect were concatenated, separated by a special [SEP] token, and then tokenized. The architecture of the Bi-LSTM included an embedding layer, followed by a two-layer bidirectional LSTM, and a fully connected layer for classification into positive, neutral, and negative sentiment classes. We used cross-entropy loss as the objective function and the Adam optimizer. In our experiments, we explored different combinations of hyperparameters, such as learning rates, batch sizes, and the number of epochs, to identify the optimal configuration. The model was trained and validated on a split dataset (70% training, 15% validation, and 15% test) to fine-tune these hyperparameters. We observed that, with a learning rate of 1×10^{-3}, a batch size of 8, and training for 10 epochs, the Bi-LSTM achieved its best performance on the test set, with an accuracy of 73.18%, a macro-precision of 71.65, recall of 69.83, and an F1 score of 70.55. This configuration was chosen as the final model for comparison against the transformer models.

When comparing the traditional Bi-LSTM model to the transformer-based models, there is a notable difference in performance. The Bi-LSTM achieved an accuracy of 72.55%, with a macro-precision of 71.16, recall of 69.98, and F1 score of 70.43, making it the lowest-performing model in this comparison. The gap between the Bi-LSTM and the transformer models is substantial, with the transformer models consistently outperforming the Bi-LSTM by significant margins across all metrics. This difference can be attributed to the transformers' ability to leverage pre-trained language understanding and capture long-range dependencies more effectively, whereas the Bi-LSTM, while capable of handling sequence data, lacks the pre-training and sophisticated attention mechanisms of transformers. This underscores the advantage of using transformer-based architectures for tasks like an aspect-based sentiment analysis, where the context and relationships between words play a critical role in model performance.

The results also indicate that GPT-4 outperforms GPT-3.5-turbo across all metrics, achieving an accuracy of 78.70% and an F1 score of 73.52, compared to GPT-3.5-turbo, which has an accuracy of 73.89% and an F1 score of 61.39. While GPT-4 demonstrates stronger performance overall, both LLMs perform significantly lower than the fine-tuned transformer models. The relatively lower F1 scores for both LLMs can be attributed to the fact that they were not fine-tuned on this dataset and were instead tested on the entire dataset using their general-purpose capabilities, whereas the transformer models were fine-tuned and tested on a 15% split, optimizing them for this specific task. This lack of fine-tuning likely impacted the LLMs' ability to handle the nuances of an aspect-based sentiment analysis (ABSA), resulting in lower recall and F1 scores. When compared to the fine-tuned transformer models like RoBERTa-base (which achieved the highest F1 score of 88.08), BERT, and XLNet, GPT-3.5 and GPT-4.0 report lower performance. Fine-tuned models like RoBERTa and BERT-large consistently outperform the LLMs, with accuracy scores close to or above 87% and F1 scores over 85%, showcasing the critical advantage of task-specific fine-tuning. When comparing the non-transformer models like Bi-LSTM, LSTM, and RNN with GPT-4 and GPT-3.5-turbo, we observe that the non-transformer models, particularly Bi-LSTM (F1 score: 70.55), perform surprisingly close to GPT-3.5-turbo, and in some cases, even surpass it in certain metrics like recall. However, GPT-4 still outperforms these non-transformer models, with an F1 score of 73.52%. The key difference here is that the non-transformer models were fine-tuned and trained on a portion of the dataset, allowing them to adapt better to the specific task, while the LLMs were tested in a zero-shot setting without fine-tuning. Despite this, the non-transformer models lack the overall versatility and adaptability of LLMs, which, with fine-tuning, would likely far surpass their performance. The results highlight that while non-transformer models can perform decently when properly trained, LLMs like GPT-4 have the potential for much greater performance, especially when fine-tuned for task-specific datasets like ABSA. However, both non-transformer models and GPT-3.5-turbo fall well below the performance of the fine-tuned transformers. This highlights that while LLMs like GPT-4 and GPT-3.5-turbo are powerful general models, they still struggle to match the performance of fine-tuned transformers in specialized tasks like ABSA without further fine-tuning.

3.4.2. Results on Naver Labs Europe Dataset

In addition to evaluating our models on the MAMS and SemEval datasets that we created, we further tested their performance on the Naver Labs Europe dataset to assess the models' ability to generalize to unseen data from a different source. In Table 2, the performance of the models on the Naver Labs dataset is illustrated.

Table 2. The performance results of the best fine-tuned models on the Naver Labs Europe Dataset. The best model is highlighted in bold.

Model	Accuracy	Precision	Recall	F1 Score
bert-base-uncased	96.03%	89.72	91.98	90.80
bert-large-uncased	95.24%	88.65	88.65	88.65
distilbert-base-uncased	96.83%	94.83	89.55	91.96
albert-base-v1	92.86%	84.87	78.65	81.33
albert-large-v1	95.24%	**97.44**	80.00	86.18
roberta-base	**97.62%**	93.30	95.77	94.48
roberta-large	**97.62%**	91.67	**98.65**	**94.77**
xlnet-base-cased	94.44%	89.47	82.43	85.48
Bi-LSTM	87.30%	67.52	61.08	63.16
LSTM	87.30%	48.46	46.49	47.45
RNN	88.10%	70.00	58.65	61.04
GPT-4	86.95%	64.07	73.81	63.94
GPT-3.5-turbo	94.21%	63.42	66.53	64.76

The results from the Naver dataset demonstrate the clear superiority of transformer-based models over traditional RNN, LSTM, and Bi-LSTM models, further solidifying the impact of recent advancements in natural language processing. Among the transformer models, RoBERTa (both base and large) stands out, achieving the highest accuracy of 97.62%. The large variant not only excels in accuracy but also demonstrates exceptional performance in the recall (98.65%) and F1 score (94.77%), indicating its ability to correctly identify and classify sentiments in a wide variety of contexts. This performance shows the robustness of RoBERTa's architecture in handling diverse datasets and reinforces its effectiveness in aspect-based sentiment analysis (ABSA) tasks, even when dealing with real-world reviews that contain multiple sentiment aspects. DistilBERT also performs remarkably well with an accuracy of 96.83%, proving that even lighter, distilled versions of transformer models can outperform larger models like BERT and ALBERT in certain contexts. DistilBERT's high precision (94.83%) and balanced F1 score (91.96%) show that, despite its reduced size, it effectively balances resource efficiency with performance, making it a valuable option for scenarios where computational efficiency is critical without sacrificing much accuracy.

On the other hand, ALBERT-base and XLNet-base-cased showed weaker performance across all metrics. ALBERT-base achieved 92.86% accuracy, but with noticeably lower precision (84.87%) and recall (78.65%), suggesting that its architecture might struggle with the specific characteristics of the Naver dataset, which includes varied review structures and nuanced sentiments. Similarly, XLNet-base-cased achieved 94.44% accuracy, but its lower F1 score (85.48%) points to potential difficulties in handling an aspect-specific sentiment, a task where RoBERTa excels due to its architecture's robust contextual understanding.

In contrast, traditional models such as Bi-LSTM, LSTM, and RNN lag significantly behind. Among them, Bi-LSTM showed the best performance, but still only reached 87.30% accuracy, coupled with much lower precision, recall, and F1 scores (63.16% F1 score). This performance gap underscores the limitations of older neural network models, particularly in capturing the contextual dependencies and nuanced sentiments found in more complex language tasks. LSTM and RNN performed even worse, further highlighting how traditional architectures struggle to match the capabilities of modern transformers when applied to tasks like ABSA.

These results not only highlight the effectiveness of transformer architectures in handling a complex, multi-aspect sentiment analysis but also demonstrate their generalization capabilities across different datasets. While most models performed reasonably well, RoBERTa continues to lead the way, showing that it is particularly well suited for cross-domain applications and tasks that require deep contextual understanding. The results reaffirm the value of transformer models in sentiment analyses, suggesting that they are better equipped to handle the nuances of the language and aspect-level sentiment in real-world applications. Moreover, the fact that even DistilBERT performs exceptionally well reinforces the practical benefits of transformer models, offering high performance without the resource-heavy requirements of larger models.

Finally, GPT-4 achieves an accuracy of 86.95% with a precision of 64.07, recall of 73.81, and F1 score of 63.94, while GPT-3.5-turbo performs slightly better in accuracy (94.21%), but with lower precision (63.42) and recall (66.53), and a marginally higher F1 score (64.76). Similar to the MAMS Semeval results, both LLMs underperform compared to fine-tuned transformer models, largely due to the fact that GPT-4 and GPT-3.5-turbo were not fine-tuned on the Naver dataset and were instead evaluated on the entire dataset using their general-purpose capabilities. This lack of fine-tuning impacts the models' performance in the sentiment analysis task, where task-specific training is crucial for higher performance. In comparison, fine-tuned transformer models such as RoBERTa-base and RoBERTa-large dominate, achieving the highest accuracy (97.62%) and F1 scores (94.48 and 94.77, respectively). Fine-tuned models like BERT and DistilBERT also perform significantly better than the LLMs, further highlighting the advantage of fine-tuning for task-specific datasets like this one. When comparing LLMs to non-transformer models, such as Bi-LSTM and RNN, the results are mixed. Bi-LSTM, with an F1 score of 63.16, performs similarly to both GPT-4 and GPT-3.5-turbo, while LSTM and RNN underperform with F1 scores of 47.45 and 61.04, respectively. Despite this, LLMs still generally outperform the non-transformer models, but fall short of the fine-tuned transformer models, underscoring the importance of fine-tuning for these specific tasks.

4. Explainability Techniques on the Transformer Models

Transformer-based models such as RoBERTa are highly effective due to their sophisticated architectures and ability to capture intricate dependencies in text. However, their complexity makes it challenging to interpret their decision-making processes. As models become more powerful, they often sacrifice transparency, making it difficult for users to fully understand how predictions are made. This presents a significant challenge when deploying these models in real-world applications where trust, fairness, and transparency are crucial. So, explainability techniques such as LIME, SHAP, and integrated gradients can help to shed light on the models' internal operations. While these techniques enhance the interpretability of complex models, they can introduce approximations, highlighting the trade-off between achieving state-of-the-art performance and maintaining a clear understanding of the model's decisions. Explainability techniques can shed light on the decision-making processes of these models, making their outputs more transparent and understandable. So, we implemented and used five explainability techniques, which are the LIME (Local Interpretable Model-agnostic Explanation) [35], SHAP (SHapley Additive exPlanation) [36], Grad-CAM (Gradient-weighted Class Activation Mapping) [37], integrated gradients [38], and Attention Visualization [39].

LIME (Local Interpretable Model-agnostic Explanation) is a model-agnostic technique that approximates the predictions of a complex model with an interpretable model, such as a linear regression or decision tree, for a specific instance. By perturbing the input and observing the changes in the output, LIME identifies the most important features contributing to the prediction. This helps us in understanding which words or phrases in a text are driving the model's sentiment classification, providing insights into how the model reaches its decisions. This technique is particularly useful as it clarifies the impact of individual textual components on the overall sentiment prediction.

The SHAP (SHapley Additive exPlanation) technique aims to explain the output of a model by distributing the prediction among the input features based on their contribution. It can highlight which words or aspects of a sentence are influencing the sentiment prediction, offering us a deeper understanding of the transformer model's behavior and the rationale behind its predictions. This technique helps us to ensure the transformer model's transparency and accountability by illustrating the contribution of each feature.

The Attention Visualization technique highlights which words in a text are attending to which other words, revealing patterns and relationships that the model is leveraging to make predictions. This can be particularly useful in understanding the context and dependencies considered by the model in aspect-based sentiment analysis tasks. By visualizing the attention weights in our transformer models, we can see how the model prioritizes certain words or phrases, providing a deeper understanding of the model's interpretive process.

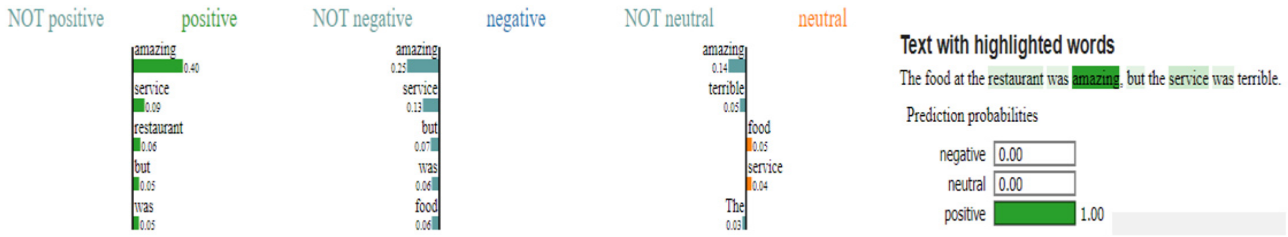
The integrated gradient technique attributes the prediction of a model to its input features by integrating the gradients of the model's output with respect to the input along a path from a baseline to the actual input. This technique provides us with a way to quantify the contribution of each word or phrase to the transformer model's prediction, offering us insights into how the transformer arrives at its particular decision. Since transformers rely heavily on self-attention mechanisms, visualizing attention weights through integrated gradients can greatly help us and reveal to us how the model prioritizes different parts of the input text, enhancing the interpretability of complex transformer models.

The Grad-CAM (Gradient-weighted Class Activation Mapping) technique is a visualization approach that uses the gradients of the target output with respect to the input embeddings and produces a coarse localization map, highlighting important regions in the input text. This technique helps in visualizing which parts of the text are most influential for the model's prediction. By applying Grad-CAM to NLP models, we can gain a better understanding of the areas within the text that significantly impact the model's outputs, thereby enhancing the transformer model's transparency and interpretability.

The implementation of these explainability techniques and their use on the transformer models we fine-tuned on an aspect-based sentiment analysis can greatly assist us in understanding how exactly the models perform their decision-making procedure. In addition, it can help to understand which specific words one model utilizes and to which degree, and also how the model pays attention to specific words and features and provides us useful insights that can also assist us in further improving the performance of the models, addressing possible biases and ensuring that the models are efficient. Below, we present in detail the way that each explainability method operates, and how they shed light on the functionality of the transformer models. All results, including the screenshot images, were obtained from our Google Collab environment.

4.1. LIME Technique

The LIME aims to fit a surrogate glass-box model around the decision space of the transformer model's prediction. LIME has been designed to be applied locally and we utilized it on the best-performing fine-tuned model, which was the roberta-base transformer, to see how the model operates, and which words or phrases were the most influential in the decision making. We illustrate the functionality of the LIME technique (as well as of all the explainability techniques) on the sentence "The food at the restaurant was amazing, but the service was terrible.", which has many aspects and different polarities. Initially, for the aspect "food", the LIME analysis is illustrated in Figure 2.



Text with highlighted words

The food at the restaurant was amazing, but the service was terrible.

Prediction probabilities

negative 0.00
neutral 0.00
positive 1.00

Figure 2. The LIME analysis on the aspect "food".

As we can see in this example sentence, our transformer model has correctly predicted the target aspect polarity as positive for the aspect "food". The prediction probabilities show a 100% confidence in the positive sentiment, with no probabilities assigned to negative or neutral sentiments. The LIME explanation highlights the word "amazing" with the highest weight (0.40), which strongly influences the positive sentiment prediction. Other words of the sentences such as "service" (0.09), "restaurant" (0.06), "but" (0.05), and "was" (0.05) also contribute but to a lower extent. Our model has inferred this positive polarity due to the significant influence of the word "amazing", correctly identifying it as the key indicator of the positive sentiment. The highlighted weights show that the model considers the context and the impactful words to make its prediction, ensuring that the positive sentiment associated with "food" is accurately captured. After that, for the aspect "service", the LIME analysis is illustrated in Figure 3.

Our transformer has correctly predicted the target aspect polarity as negative for the aspect "service". The prediction probabilities show a 100% confidence in the negative sentiment, with no probabilities assigned to neutral or positive sentiments. The LIME explanation highlights the word "service" with the highest weight (0.31), which strongly influences the negative sentiment prediction. Other significant contributing words include "but" (0.29), "terrible" (0.22), and "food" (0.11). The model has inferred this negative polarity due to the strong influence of the word "service" particularly in the negative context provided by the word "terrible". The highlighted weights show that the model accurately captures the negative sentiment associated with the aspect "service" with "but" also playing a key role in emphasizing the contrast between the positive and negative parts of the sentence. The presence of "amazing" has a minimal impact on this aspect, reinforcing the model's focus on the relevant context. Finally, for the aspect "restaurant", the LIME analysis is illustrated in Figure 4.

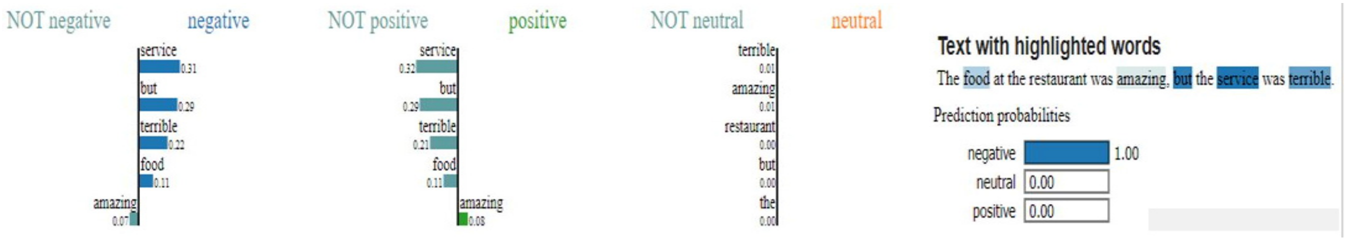


Figure 3. The LIME analysis on the aspect "service".

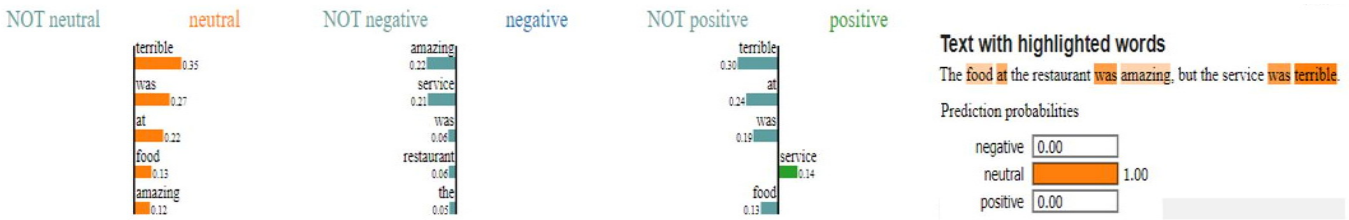


Figure 4. The LIME analysis on the aspect "restaurant".

Once again, our fine-tuned transformer model has indeed correctly predicted the target aspect polarity as neutral for the aspect "restaurant". The prediction probabilities show a 100% confidence in the neutral sentiment, with no probabilities assigned to negative or positive sentiments. The LIME explanation highlights the word "terrible" with the highest weight (0.35), which influences the sentiment prediction. Other significant contributing words include "was" (0.27), "at" (0.22), "food" (0.13), and "amazing" (0.12). The model has inferred this neutral polarity due to the balanced influence of both positive and negative words in the context of the aspect "restaurant". The word "terrible" has a high weight, indicating its strong negative sentiment, but it is counterbalanced by words like "amazing" and "food" that contribute positively. This balance results in an overall neutral sentiment for the aspect "restaurant", demonstrating the model's capability to capture the mixed sentiments present in the sentence.

4.2. SHAP Technique

The SHAP technique aims to illustrate the output of our transformer model using Shapley values. Similarly, here, we used the same transformer model and we examine once again the sentence "The food at the restaurant was amazing, but the service was terrible.", to see how the model can infer the polarity of the target aspects. Initially, for the aspect "food", the SHAP analysis is illustrated in Figure 5.

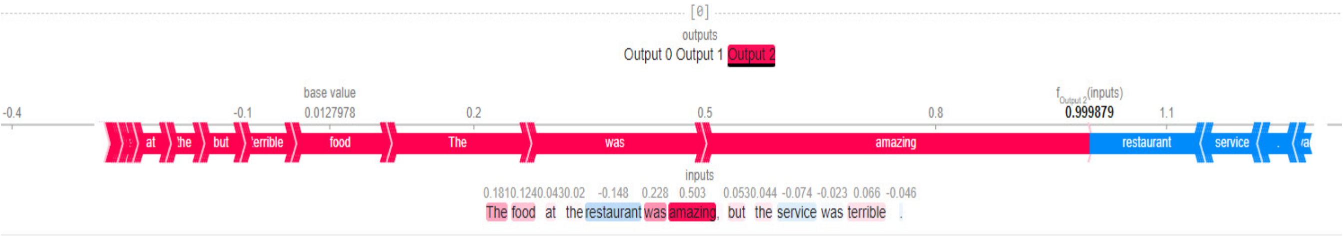


Figure 5. The SHAP analysis on the aspect "food".

SHAP results on the sentence "The food at the restaurant was amazing, but the service was terrible" indicate that our fine-tuned transformer model has correctly predicted the target aspect polarity for "food" as positive (Output 2). This inference is supported by the SHAP values, which show the weights assigned to different words in the sentence. The word "amazing" has a significant positive weight (0.503), strongly influencing the positive sentiment prediction. Although the word "terrible" has a value of 0.06, it does not significantly impact the sentiment prediction for the aspect "food". The model has effectively focused on the relevant parts of the sentence related to the food aspect, leading to an accurate sentiment prediction. After that, for the aspect "service", the SHAP analysis is depicted in Figure 6.

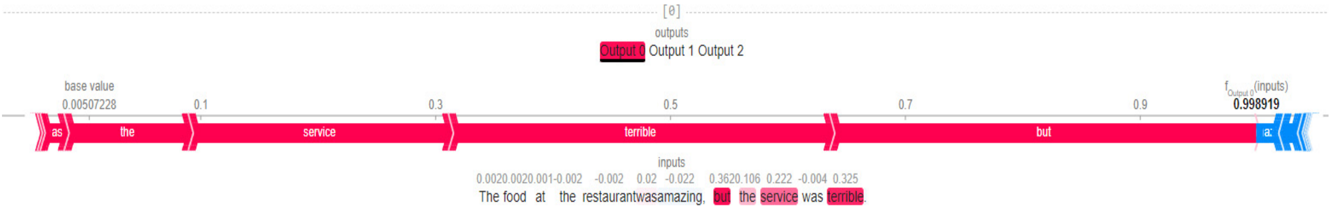


Figure 6. The SHAP analysis on the aspect "service".

We can see that our fine-tuned transformer model has correctly predicted the target aspect polarity for "service" as negative (Output 0). This inference is supported by the SHAP values, which show the weights assigned to different words in the sentence. The words "but" (0.3620) and "terrible" (0.325) have significant positive weights, strongly influencing the negative sentiment prediction. The word "service" (0.222) also contributes to the overall sentiment analysis, indicating a contrast in the sentence structure. The model has effectively focused on the relevant parts of the sentence related to the service aspect, leading to an accurate sentiment prediction. Finally, for the aspect "restaurant", the SHAP analysis is illustrated in Figure 7.

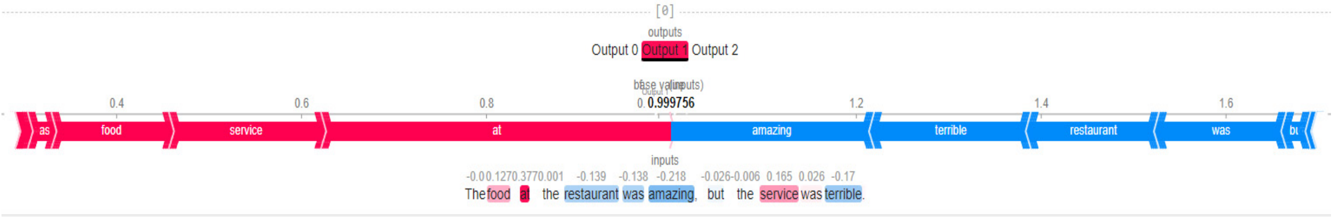


Figure 7. The SHAP analysis on the aspect "restaurant".

It can be observed that our fine-tuned transformer model has correctly predicted the target aspect polarity for "restaurant" as neutral (Output 1). This inference is supported by the SHAP values, which show the weights assigned to different words in the sentence. For the aspect "restaurant", the word "at" has a significant positive weight of 0.377, contributing positively towards the prediction of a neutral sentiment. Similarly, "food" has a positive weight of 0.127, and "service" has a positive weight of 0.165. These words positively influence the sentiment prediction towards neutrality. The word "amazing" has a negative weight, which would typically pull the prediction towards a positive sentiment. Conversely, the word "terrible" also has a negative weight, pulling towards a negative sentiment. However, the model balances these influences, with the contributions from "at", "food", and "service" leading to a neutral overall sentiment for the aspect "restaurant". This balanced interpretation by the model results in the correct neutral prediction for the aspect "restaurant".

4.3. Attention Weight Visualization

The attention weight visualization technique we implemented and used in our model highlights the parts of the input text that our model focuses on when making a prediction. So, in this way, the attention technique can provide insights on which words or phrases are most influential in determining the sentiment of a particular aspect. The attention weight visualization technique was applied on our fine-tuned RoBerta model for the same sentence, "The food at the restaurant was amazing, but the service was terrible.", and initially, for the aspect "food", the attention weights are illustrated in Figure 8.

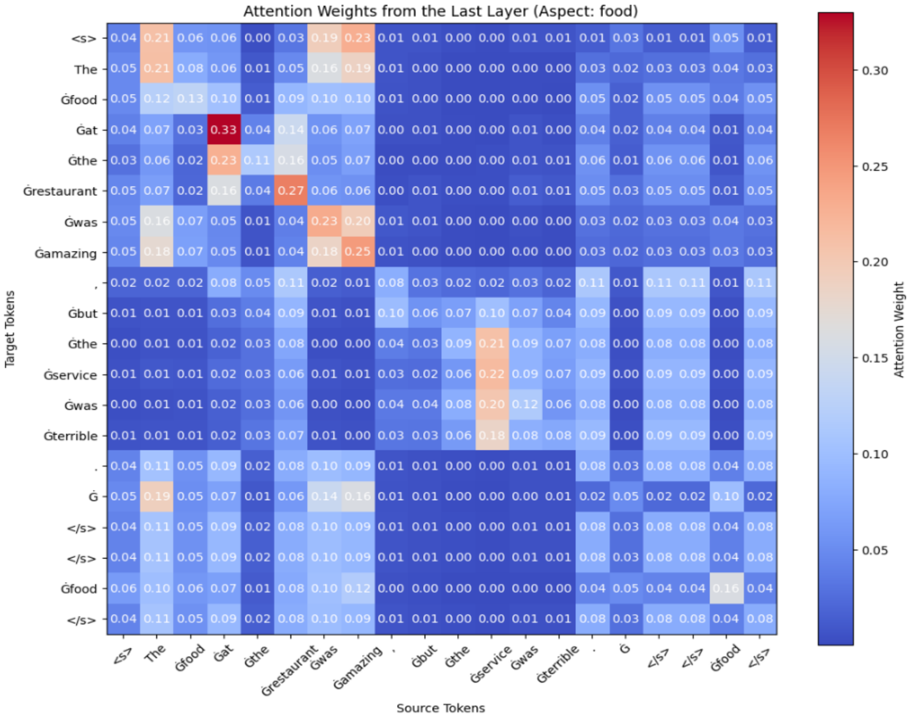


Figure 8. The visualization of the attention weights for the aspect "food".

The attention weights that our fine-tuned transformer created by focusing on the aspect word "food" become clear. The attention weights indicate how much attention the model is assigning from each source token to each target token. Based on the plot, it is noted that the same word in the source has a high attention weight with the same word in the target, which indicates that the model is paying close attention to the self-referential information within the text. Furthermore, there are high attention weights between tokens such as "." and "<s>" and the target tokens, which indicates that the model focuses on understanding the structure and boundaries of the input sequence, crucial for the accurate processing and generation of text. Specifically, here, the word "food" in the source has a high attention weight with the word "amazing" (0.07) in the target, as well as the word "amazing" in the source having a high attention weight with the word "food" (0.10) in the target, which shows that the model is focusing on the relationship between these two words. In addition, "food" has a relatively low attention weight with the word "terrible" (0.01) in the target as well as the word "terrible" in the source having zero attention weight with the word "food" (0.00) in the target. This indicates that the model has correctly focused on the contextual content of the aspect word "food" (the words that are related to the aspect word) and ignores the rest of the words related to other aspects. In Figure 9, the attention weights for the aspect "service" are illustrated.

In this figure, we can observe the attention weights that our fine-tuned transformer created by focusing on the aspect word "service". Specifically, in this case, the word "service" has a high attention weight with the word "terrible" (0.17) in the target, as well as the word "terrible" in the source having a high attention weight with the word "service"

(0.19) in the target, which shows that the model is focusing on the relationship between these two words. Additionally, "service" has a significantly low attention weight with the word "amazing" (0.01) in the target, and "amazing" in the source has zero attention weight with "service" (0.00) in the target. This indicates that the model has correctly focused on the contextual content of the aspect word "service" and ignores the rest of the words related to other aspects. In Figure 10, the attention weights for the aspect "restaurant" are illustrated.

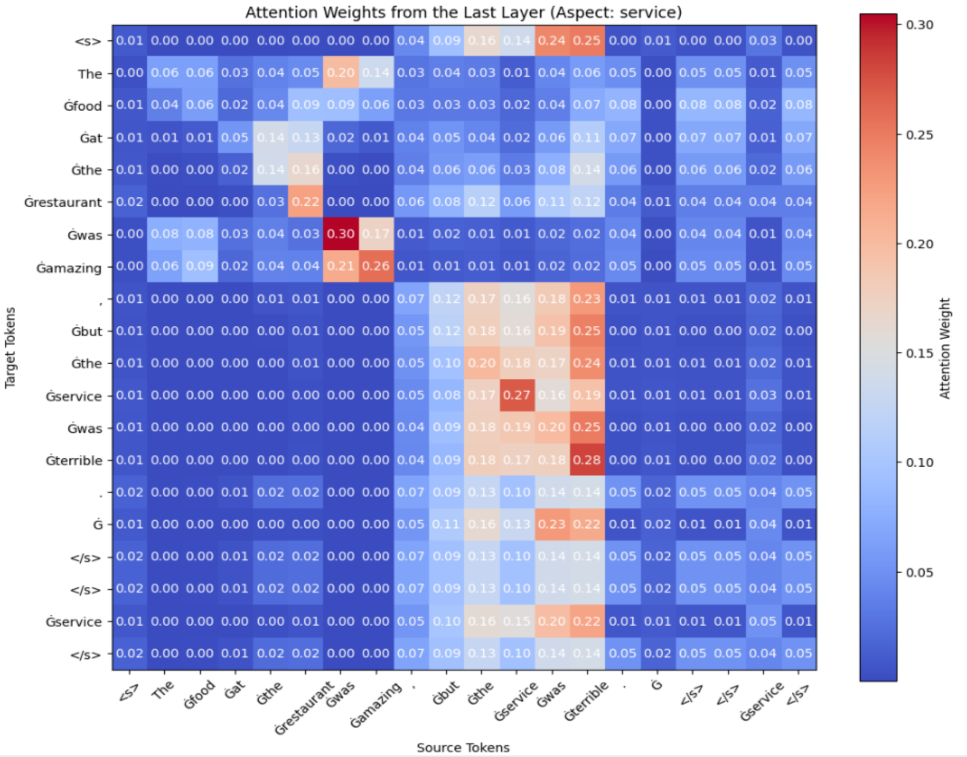


Figure 9. The visualization of the attention weights for the aspect "service".

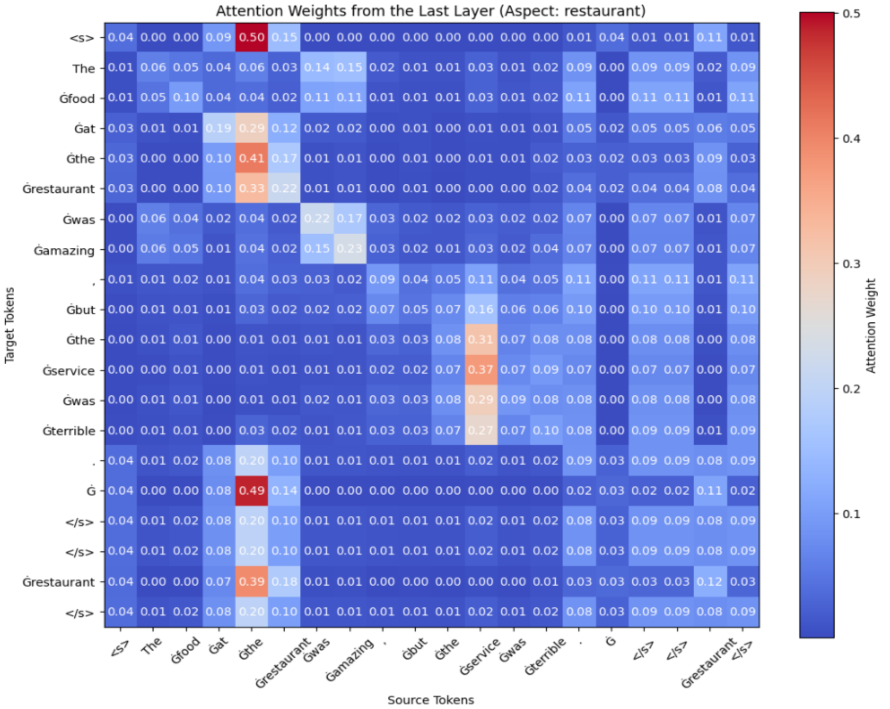


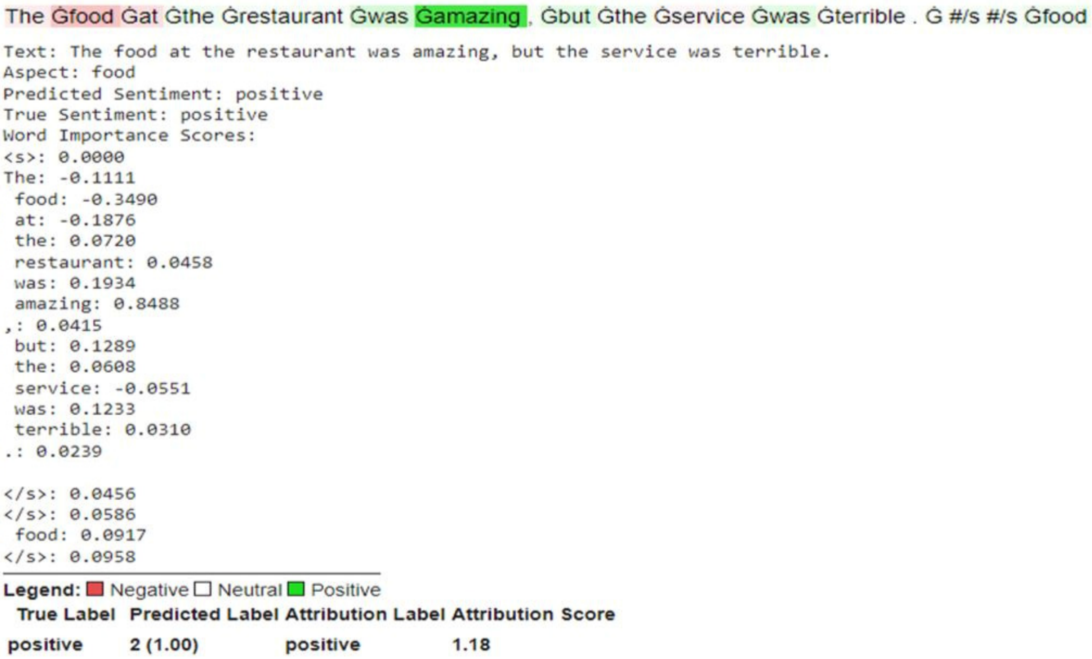
Figure 10. The visualization of the attention weights for the aspect "restaurant".

In this figure, we can examine the attention weights that our fine-tuned transformer created by focusing on the aspect word "restaurant". In this case, the word "restaurant" in the source has a 0.02 attention weight with the word "amazing" in the target, as well as the word "amazing" in the source having a weight of 0.01 with the word "restaurant" in the target. Additionally, "restaurant" has a 0.02 attention weight with the word "terrible" in the target, and "terrible" in the source has a 0.02 attention weight with "restaurant" in the target. It is obvious that our model, focusing on the aspect word "restaurant", has given the same attention weight, 0.02, for the words "amazing" and "terrible", implying that both words participate equally in the prediction of polarity.

4.4. Integrated Gradients

The integrated gradient technique was applied on our transformer model and aims to attribute an importance value to each input feature of the model based on the gradients of the model output with respect to the input. The technique can provide quite valuable insights into how each word in the input text contributes to the final prediction. By integrating the gradients along the path from a baseline input (typically a neutral or zero input) to the actual input, this method effectively highlights the contribution of each feature. This assists in understanding which aspects of the input text are the most influential in determining the model's sentiment classification.

Similarly, we used our RoBERTa fine-tuned model and we examined the sentence "The food at the restaurant was amazing, but the service was terrible". For the aspect "food", the integrated gradients are illustrated in Figure 11.



The Ġfood Ġat Ġthe Ġrestaurant Ġwas Ġamazing , Ġbut Ġthe Ġservice Ġwas Ġterrible . Ġ #/s #/s Ġfood

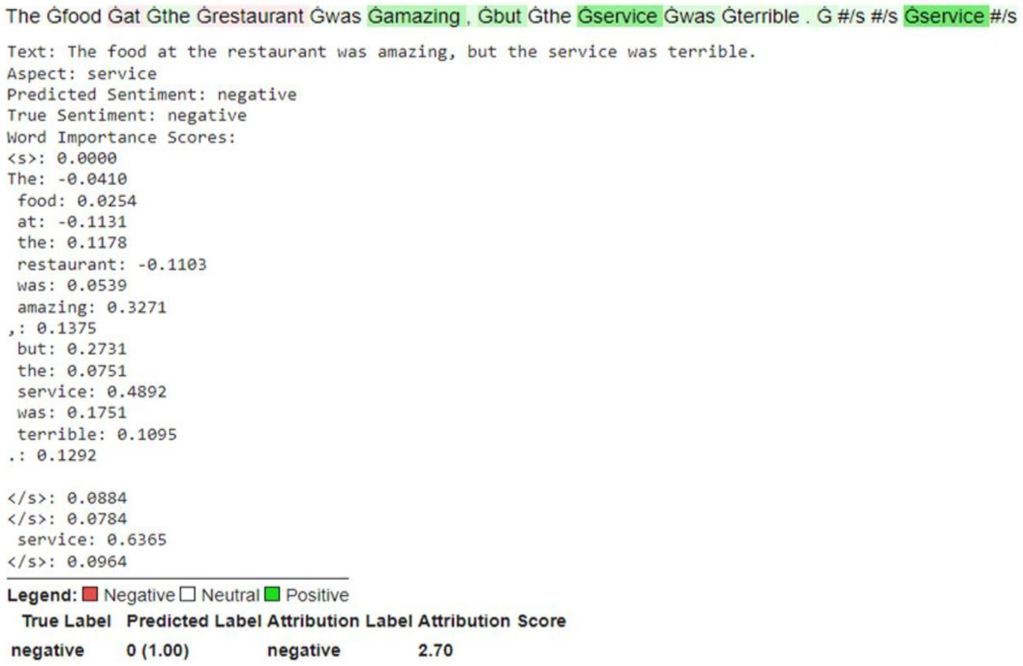
```
Text: The food at the restaurant was amazing, but the service was terrible.
Aspect: food
Predicted Sentiment: positive
True Sentiment: positive
Word Importance Scores:
<s>: 0.0000
The: -0.1111
 food: -0.3490
 at: -0.1876
 the: 0.0720
 restaurant: 0.0458
 was: 0.1934
 amazing: 0.8488
,: 0.0415
 but: 0.1289
 the: 0.0608
 service: -0.0551
 was: 0.1233
 terrible: 0.0310
.: 0.0239

</s>: 0.0456
</s>: 0.0586
 food: 0.0917
</s>: 0.0958
```

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score
positive	2 (1.00)	positive	1.18

Figure 11. Integrated gradients for the aspect "food".

It is noted that our model has correctly predicted the target aspect polarity as positive for the aspect "food" and has inferred, e.g., the weights assigned to specific words. The word "amazing" has the highest positive weight (0.8488), strongly contributing to the positive sentiment. Despite some negative weights, such as for "food" (−0.3490) and "service" (−0.0551), the overall positive sentiment is primarily driven by the highly positive weight for "amazing". This indicates that the model effectively focuses on the most relevant words that convey the sentiment for the aspect "food". Thereafter, the integrated gradients for the aspect "service" are illustrated in Figure 12.



The Ġfood Ġat Ġthe Ġrestaurant Ġwas Ġamazing , Ġbut Ġthe Ġservice Ġwas Ġterrible . Ġ #/s #/s Ġservice #/s

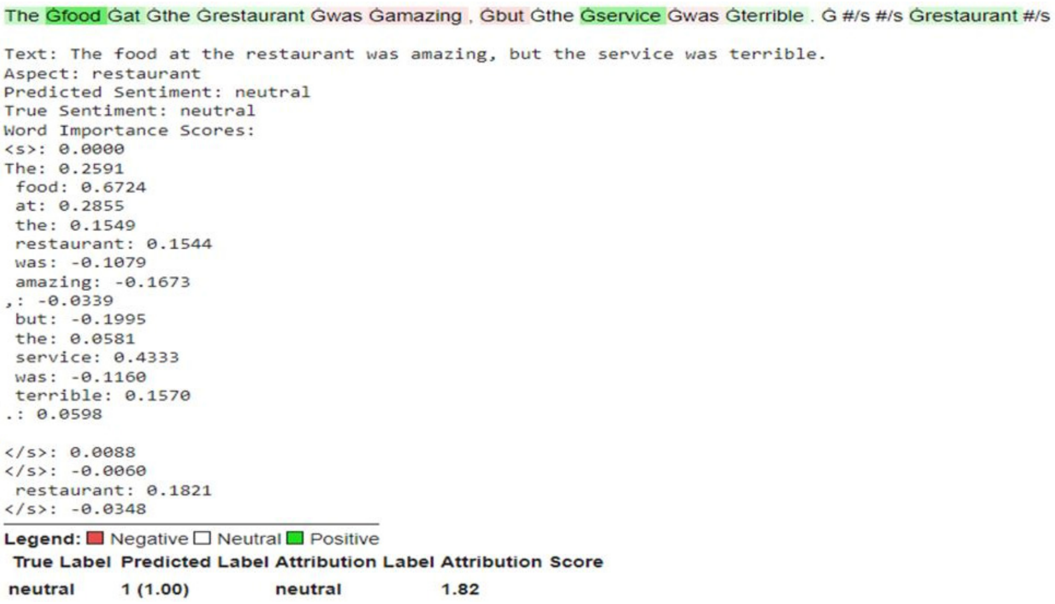
```
Text: The food at the restaurant was amazing, but the service was terrible.
Aspect: service
Predicted Sentiment: negative
True Sentiment: negative
Word Importance Scores:
<s>: 0.0000
The: -0.0410
 food: 0.0254
 at: -0.1131
 the: 0.1178
 restaurant: -0.1103
 was: 0.0539
 amazing: 0.3271
,: 0.1375
 but: 0.2731
 the: 0.0751
 service: 0.4892
 was: 0.1751
 terrible: 0.1095
.: 0.1292

</s>: 0.0884
</s>: 0.0784
 service: 0.6365
</s>: 0.0964
```

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score
negative	0 (1.00)	negative	2.70

Figure 12. Integrated gradients for the aspect "service".

The model has correctly predicted the target aspect polarity as negative. It has inferred this by focusing on the word "service", which has high attribution scores (0.4892 and 0.6365), indicating its crucial role in the model's prediction. These scores do not suggest that "service" itself contributes positively to a negative sentiment but rather highlight its importance in the context of the sentence. The model considers the negative connotation of "terrible" (0.1095) directly associated with "service" to make the prediction. Additionally, the conjunction "but" (0.2731) signals a contrast, emphasizing the negative sentiment towards "service" despite the positive sentiment for "food". The high attribution for "amazing" (0.3271) further underscores this contrast. Thus, the model captures the context and the relationships between words to accurately predict a negative sentiment for the aspect "service". Finally, for the aspect "restaurant", the integrated gradients are illustrated in Figure 13.



The Ġfood Ġat Ġthe Ġrestaurant Ġwas Ġamazing , Ġbut Ġthe Ġservice Ġwas Ġterrible . Ġ #/s #/s Ġrestaurant #/s

```
Text: The food at the restaurant was amazing, but the service was terrible.
Aspect: restaurant
Predicted Sentiment: neutral
True Sentiment: neutral
Word Importance Scores:
<s>: 0.0000
The: 0.2591
 food: 0.6724
 at: 0.2855
 the: 0.1549
 restaurant: 0.1544
 was: -0.1079
 amazing: -0.1673
,: -0.0339
 but: -0.1995
 the: 0.0581
 service: 0.4333
 was: -0.1160
 terrible: 0.1570
.: 0.0598

</s>: 0.0088
</s>: -0.0060
 restaurant: 0.1821
</s>: -0.0348
```

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score
neutral	1 (1.00)	neutral	1.82

Figure 13. Integrated gradients for the aspect "restaurant".

Our model has predicted correctly the target aspect polarity as neutral. The model has inferred this by focusing on the word “restaurant”, which has positive attribution scores (0.1544 and 0.1821), indicating its crucial role in the model’s prediction. These scores do not suggest that “restaurant” itself contributes positively or negatively to the sentiment but rather highlight its importance in the context of the sentence. The word “food” has a high positive attribution score (0.6724), suggesting it strongly influences the sentiment about the “restaurant”. Similarly, “service” (0.4333) and “terrible” (0.1570) have positive attribution scores, indicating their significant influence on the sentiment. The word “amazing” has a negative score (−0.1673), which could be due to its contribution to a mixed sentiment context rather than directly influencing the restaurant’s sentiment negatively. Moreover, the conjunction “but” (−0.1995) indicates a contrast, helping the model understand the nuanced sentiment. Thus, the model captures the context and the relationships between words to accurately predict a neutral sentiment for the aspect “restaurant”.

4.5. Grad-CAM Technique

We also implemented and used the Grad-CAM (Gradient-weighted Class Activation Mapping) on our fine-tuned model. Grad-CAM identifies and visualizes the relevance of each word in the text to the sentiment prediction for a given aspect. First, a forward pass obtains the transformer model’s predictions and activations from a selected layer. Following this, during the backward pass, gradients of the target sentiment score with respect to these activations are computed. These gradients, averaged and weighted, highlight the importance of each token in the input text. The resulting relevance scores are then visualized, often as a heatmap, showing which parts of the text the model considers most influential in determining the sentiment towards the specified aspect. This process helps us in interpreting and understanding the model’s decision making for specific aspects in the text. Similarly, we used our RoBERTa fine-tuned model and we examined the sentence “The food at the restaurant was amazing, but the service was terrible”. For the aspect “food”, the results of the Grad-CAM are illustrated in Figure 14.

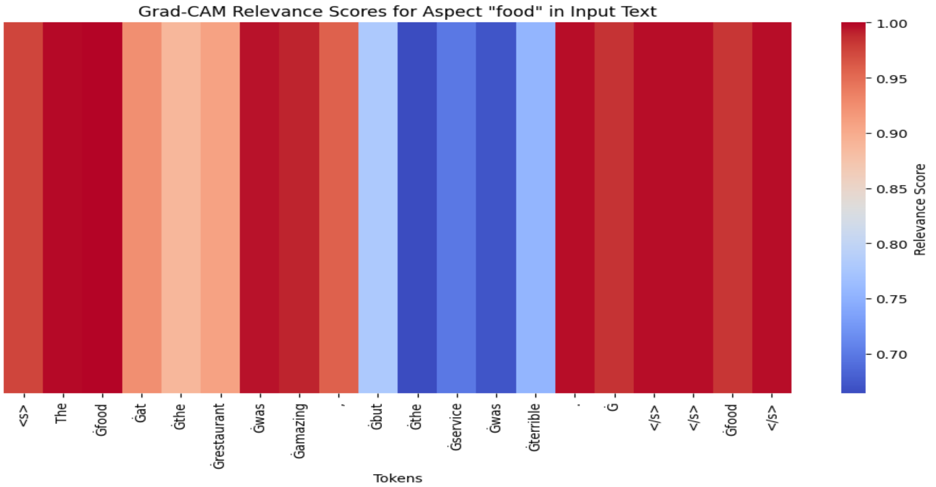


Figure 14. Grad-CAM for the aspect “food”.

The Grad-CAM relevance scores indicate that the model has heavily weighted certain words when making its prediction. For instance, the word “food” has the highest relevance score of 1.0000, indicating that it had the most significant impact on the model’s decision. This makes sense, as “food” is the main aspect being evaluated. Other words like “The” (0.9981), “was” (0.9951), and “amazing” (0.9905) and punctuation marks like “.” (0.9969) also have high relevance scores, showing their influence on the model’s decision. Interestingly, words related to the negative part of the sentence, such as “service” (0.6991), “was” (0.6707), and “terrible” (0.7544), have relatively lower relevance scores compared to positive words. This suggests that while the model considered the entire sentence, it weighed the positive sentiment associated with “food” more heavily. Overall, the model’s prediction aligns well with the expected sentiment for the aspect “food”, correctly focusing on the relevant positive indicators in the text. After that, for the aspect “service”, the Grad-CAM analysis is illustrated in Figure 15.

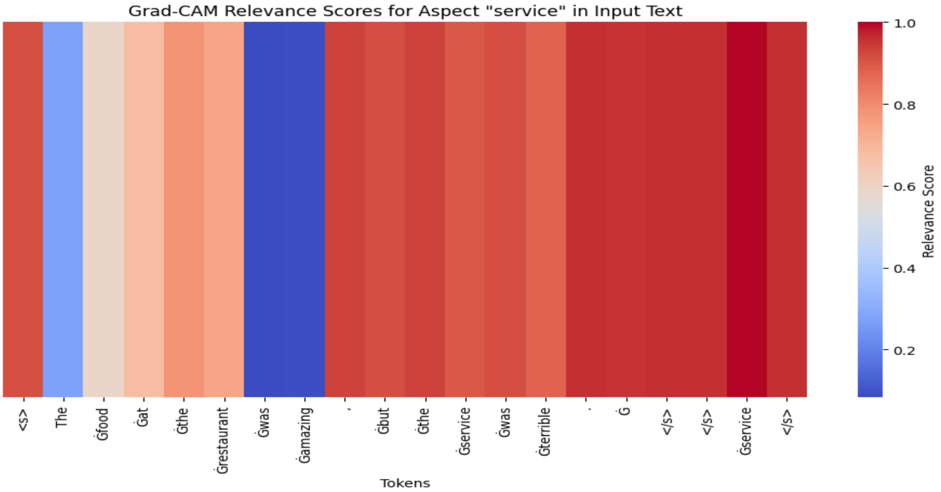


Figure 15. Grad-CAM for the aspect “service”.

The model has correctly predicted the target aspect polarity as negative for the aspect “service”. The Grad-CAM relevance scores indicate that the model has heavily weighted certain words when making its prediction. For instance, the word “service” has the highest relevance score of 1.0000, indicating that it had the most significant impact on the model’s decision. This makes sense, as “service” is the main aspect being evaluated. Other words like “terrible” (0.8792) and “was” (0.9152) and punctuation marks like “.” (0.9605) also have high relevance scores, showing their influence on the model’s decision. The words related to the positive part of the sentence, such as “food” (0.5906), “amazing” (0.0901), and “was” (0.0830), have relatively lower relevance scores compared to the negative words. This suggests that while the model considered the entire sentence, it weighed the negative sentiment associated with “service” more heavily. Additionally, words that signal a contrast, such as “but” (0.9150), also have high relevance scores, indicating that the model correctly identified the shift in the sentiment within the sentence. Overall, the model’s prediction aligns well with the expected sentiment for the aspect “service”, correctly focusing on the relevant negative indicators in the text. Finally, for the restaurant aspect, the analysis of the Grad-CAM is illustrated in Figure 16.

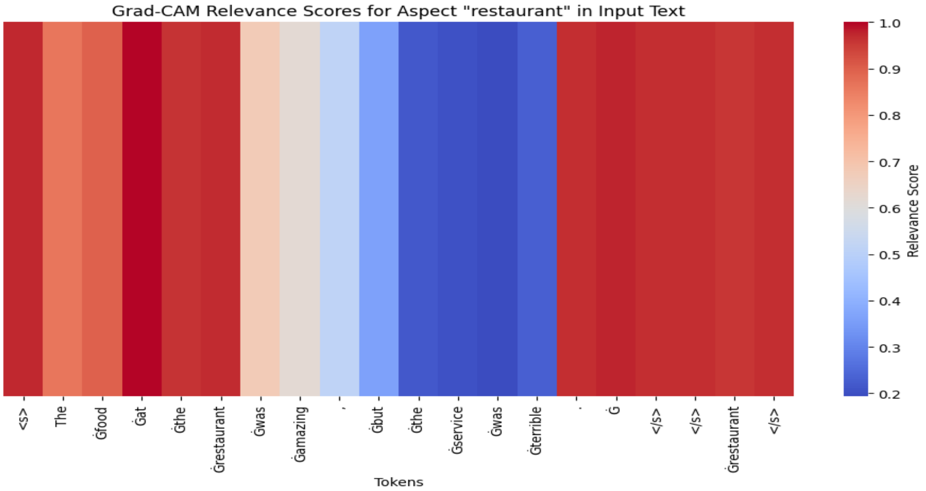


Figure 16. Grad-CAM for the aspect “restaurant”.

The Grad-CAM relevance scores indicate that the model has heavily weighted certain words when making its prediction. For instance, the word "at" has the highest relevance score of 1.0000, indicating that it had a significant impact on the model's decision. Other highly relevant words include "restaurant" (0.9704), "The" (0.8617), and "food" (0.8970), showing their influence on the model's decision. These words together help to frame the aspect "restaurant" within the context of the sentence. Interestingly, the words that express a strong sentiment, both positive and negative, such as "amazing" (0.6158), "service" (0.2040), and "terrible" (0.2298), have relatively lower relevance scores. This suggests that while the model considered the entire sentence, it weighed the contextual words around "restaurant" more heavily, which are more neutral in the sentiment. The model also gave lower relevance scores to the contrasting conjunction "but" (0.3617) and other less significant words like "was" (0.1932), and "the" (0.2173), which further indicates its focus on the main aspect-related words to determine the sentiment. Overall, the model's prediction aligns well with the expected sentiment for the aspect "restaurant", correctly focusing on the neutral context around the aspect "restaurant" in the text.

5. Discussion

The model predicted a positive sentiment for the aspect "food" in the sentence "The food at the restaurant was amazing, but the service was terrible.". Comparing the results from each explainability method provides a comprehensive understanding of how the model arrived at this conclusion. LIME shows a 100% confidence in the positive sentiment, with "amazing" having the highest weight (0.40), indicating its strong influence on the positive sentiment prediction. Other words like "service" (0.09) and "restaurant" (0.06) contribute to a lesser extent. SHAP also highlights "amazing" with a significant positive weight (0.503), reinforcing its influence on the positive sentiment prediction, while "terrible" has a minimal impact (0.06). The visualization of attention weights shows a high focus on the relationship between "food" and "amazing" (source "food"–target "amazing": 0.07 and source "amazing"–target "food": 0.10), with minimal attention to "terrible" ("food"–"terrible": 0.01 and "terrible"–"food": 0.00), confirming the model's emphasis on the positive relationship. Integrated gradients assign the highest positive weight to "amazing" (0.8488), which is significantly higher than LIME and SHAP, despite some negative weights for "food" (-0.3490) and "service" (-0.0551), indicating a strong influence of "amazing" on the positive sentiment. Grad-CAM relevance scores show "food" with the highest score (1.0000), followed by "amazing" (0.9905), and lower scores for negative words like "service" (0.6991) and "terrible" (0.7544), indicating a strong focus on the positive sentiment associated with "food". Comparing LIME and SHAP, both methods highlight "amazing" as the key driver of the positive sentiment, with SHAP assigning a slightly higher weight to "amazing". The visualization of attention weights and Grad-CAM both emphasize the relationship and relevance of "food" and "amazing", confirming the positive sentiment. Integrated gradients show the highest influence of "amazing", aligning with the findings of LIME and SHAP. In conclusion, all methods agree that "amazing" is the primary driver of the positive sentiment for "food" with LIME and SHAP emphasizing its importance, and the visualization of attention weights, integrated gradients, and Grad-CAM confirming its strong influence and relevance in the prediction.

The model predicted a negative sentiment for the aspect "service" in the sentence "The food at the restaurant was amazing, but the service was terrible.". Comparing the results from each explainability method provides a detailed understanding of how the model arrived at this conclusion. LIME shows a 100% confidence in the negative sentiment, with "service" having the highest weight (0.31), followed by "but" (0.29) and "terrible" (0.22). This indicates that "service" and the conjunction "but" play significant roles in the negative sentiment prediction. SHAP values also highlight "but" (0.3620) and "terrible" (0.325) as having significant positive weights, with "service" (0.222) contributing to the negative sentiment. Compared to LIME, SHAP assigns higher weights to "but" and "terrible", indicating a broader context and emphasizing the contrast in the sentence. The visualization of attention weights shows a high focus on the relationship between “service” and “terrible” (source “service”–target “terrible”: 0.17 and source “terrible”–target “service”: 0.19), with minimal attention to “amazing” (source “service”–target “amazing”: 0.01 and source “amazing”–target “service”: 0.00). This method highlights the direct relationship between “service” and “terrible”, confirming the model’s focus on the negative context. Integrated gradients assign high attribution scores to “service” (0.4892 and 0.6365), with contributions from “terrible” (0.1095) and “but” (0.2731). Although the conjunction “but” has a lower score compared to LIME and SHAP, integrated gradients emphasize the importance of “service” and the negative connotation of “terrible”. Grad-CAM relevance scores show “service” with the highest score (1.0000), followed by “terrible” (0.8792) and “but” (0.9150). Compared to LIME and SHAP, Grad-CAM assigns the highest relevance to “service”, confirming its central role in the negative sentiment prediction. In summary, all methods agree that “service” and “terrible” are key drivers of the negative sentiment, with LIME and SHAP highlighting the conjunction “but” more strongly than integrated gradients. The visualization of attention weights and Grad-CAM emphasize the relationship and relevance of “service” and “terrible”, confirming the negative sentiment prediction.

The model predicted a neutral sentiment for the aspect “restaurant” in the sentence “The food at the restaurant was amazing, but the service was terrible.”. Comparing the results from each explainability method provides a nuanced view of how the model reached this conclusion. LIME shows a 100% confidence in the neutral sentiment, highlighting “terrible” with the highest weight (0.35), followed by “was” (0.27), “at” (0.22), “food” (0.13), and “amazing” (0.12). This indicates that the balance of positive and negative words around “restaurant” leads to a neutral sentiment. SHAP supports this by showing mixed contributions, with “at” (0.377) not directly affecting the sentiment of “restaurant” while “food” (0.127) and “service” (0.165) add positive weights and “amazing” and “terrible” contribute negatively. SHAP shows a more detailed balance of influences compared to LIME, emphasizing the mixed sentiment context. The visualization of attention weights reveals that the word “restaurant” has equal attention weights with “amazing” (0.02) and “terrible” (0.02), indicating that both positive and negative sentiments are considered equally. This method highlights the equal contribution of positive and negative words more clearly than LIME and SHAP. Integrated gradients assign positive scores to “restaurant” (0.1544 and 0.1821), with significant contributions from “food” (0.6724), “service” (0.4333), and “terrible” (0.1570), while “amazing” has a negative score (−0.1673). This suggests a mixed sentiment context rather than a clear positive or negative influence. Grad-CAM relevance scores show “at” (1.0000) and “restaurant” (0.9704) with high relevance, while “amazing” (0.6158), “service” (0.2040), and “terrible” (0.2298) have lower relevance. This indicates that the model focuses more on contextual words around “restaurant” rather than strong sentiment words. Comparing LIME and SHAP, both methods highlight a mix of positive and negative contributions, with SHAP providing a more detailed balance. The visualization of attention weights and integrated gradients emphasize the equal contributions of positive and negative words, while Grad-CAM focuses on the contextual words around “restaurant”. In conclusion, all methods agree on a balanced influence of positive and negative words for the neutral sentiment of “restaurant” with LIME and SHAP showing mixed contributions, the visualization of attention weights and integrated gradients highlighting equal contributions, and Grad-CAM emphasizing contextual words. All in all, summing up the application of the explainability techniques, we can consider a sentence with mixed sentiments such as “The food was excellent, but the service was slow”, where the transformer model predicted a positive sentiment for the aspect “food” and a negative sentiment for the aspect “service”. LIME highlighted the words “excellent” and “slow” as the most influential for their respective aspect predictions, providing a clear explanation of how the model arrived at its decision. Similarly, SHAP values emphasized the contributions of keywords like “excellent” (positive) and “slow” (negative) in driving the sentiment predictions. Integrated gradients further reinforced this by showing a strong attribution of the word “excellent” to the positive prediction for “food” and “slow” to the negative prediction for "service". These techniques not only confirmed the correctness of the predictions but also shed light on the internal decision-making process of the model, ensuring transparency and trustworthiness. By applying these explainability methods, we gained valuable insights into how the model interprets and processes the sentiment at the aspect level, providing users with a clearer understanding of the model's behavior.

6. Conclusions

In this study, we explored the application of transformer models to an aspect-based sentiment analysis, focusing on their performance and interpretability. We fine-tuned several pre-trained transformers, including BERT, ALBERT, RoBERTa, DistilBERT, and XLNet, on a challenging dataset we formulated based on MAMS and SemEval datasets. Each instance in the dataset consists of at least two aspects and corresponding polarities. Among the transformers, RoBERTa achieved the highest accuracy of 89.16%, showcasing its suitability in handling the complexities of an aspect-based sentiment analysis.

To enhance the transparency and understanding of these fine-tuned models, we implemented five explainability techniques: LIME, SHAP, attention weight visualization, integrated gradients, and Grad-CAM. These techniques provided valuable insights into the decision-making processes of the transformers, highlighting the influential words and phrases that significantly impact their predictions. The application of explainability techniques such as LIME, SHAP, and integrated gradients provided valuable insights into how these models make predictions by highlighting the most influential words and phrases. This not only enhances the interpretability of the models but also allows for more informed adjustments to improve their performance and mitigate potential biases. LIME and SHAP helped to decompose the model's decisions by approximating and distributing contributions among input features, thereby elucidating which parts of the text were the most influential in the sentiment prediction. Attention weight visualization offered a visual representation of which words the model focused on, revealing patterns and relationships that are crucial for the model's understanding of context. Integrated gradients quantified the contribution of each word to the model's prediction by integrating gradients from a baseline to the actual input, showing how the model attributed importance across different parts of the text. Grad-CAM provided a coarse localization map of the important regions in the text, highlighting areas that significantly impacted the model's decisions.

The insights gained from these explainability techniques not only improve our understanding of how transformer models operate but also guide us in refining these models for better performance. By addressing potential biases and ensuring the models' efficiency and reliability, we can develop more accurate and trustworthy ABSA systems. Furthermore, the detailed analysis offered by explainability techniques can help in identifying specific strengths and weaknesses of the models, providing actionable insights for model improvement and adaptation and ensuring that the models are robust and reliable. The findings indicate that combining state-of-the-art transformers with explainability methods can lead to more robust and transparent sentiment analyses, which is essential for real-world applications like customer feedback analyses and market research. These results underscore the importance of explainability in ensuring that sentiment models are not only accurate but also trustworthy and transparent in their decision-making processes.

While our study demonstrates the strengths of fine-tuned transformer-based models, particularly RoBERTa, in handling an aspect-based sentiment analysis (ABSA), it is important to acknowledge some limitations. One potential issue is the bias within the datasets used, which may not capture the full range of sentiment expressions across different domains, languages, or cultural contexts. This could limit the generalizability of our models when applied to new, unseen data. Additionally, while the models performed well in our experiments, their effectiveness may vary when used in more complex or niche domains that require further fine-tuning or domain-specific adjustments. In this context, future work will focus on addressing these limitations by experimenting with more diverse datasets, conducting domain adaptation studies, and applying additional techniques to mitigate dataset biases. This will ensure that the models are not only accurate but also adaptable and fair across different real-world applications. In addition, we plan to investigate the integration of additional and more advanced transformer architectures, such as GPT or T5, which could provide improvements in aspect identification and sentiment classification. Additionally, we aim to explore hybrid explainability techniques that combine gradient-based methods with perturbation-based approaches to offer more comprehensive and robust explanations. Incorporating domain-specific knowledge into the model training and explainability process may also improve model accuracy and interpretability. Another interesting direction for future work would be to focus on developing methods to handle implicit aspects and sentiments, which are currently challenging for most ABSA models. Finally, future research will aim at further analyzing bias in combination with explainability methods, ensuring that explainability highlights any potential biases in predictions and enhancing the overall fairness and transparency of sentiment analysis models. This constitutes the main direction that future work could focus on.

Author Contributions: Conceptualization, I.P.; methodology, I.P.; software, I.P. and A.D.; validation, I.P. and A.D.; formal analysis, I.P.; investigation, I.P. and A.D.; data curation, A.D.; writing—original draft preparation, I.P. and A.D.; visualization, I.P. and A.D.; supervision, I.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new datasets were created in this study. All code and datasets used in this study can be found in our GitHub repository: https://github.com/ThanosDiam/ABSA-Analysis-with-Explainability-Methods, accessed on 30 July 2024 (GNU General Public License v3.0.).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, B. *Sentiment Analysis and Opinion Mining*; Springer Nature: Berlin, Germany, 2022. [CrossRef]
2. Rodríguez-Ibánez, M.; Casánez-Ventura, A.; Castejón-Mateos, F.; Cuenca-Jiménez, P.M. A review on sentiment analysis from social media platforms. *Expert Syst. Appl.* **2023**, *223*, 119862. [CrossRef]
3. Zhang, W.; Li, X.; Deng, Y.; Bing, L.; Lam, W. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 11019–11038. [CrossRef]
4. Chauhan, G.S.; Nahta, R.; Meena, Y.K.; Gopalani, D. Aspect based sentiment analysis using deep learning approaches: A survey. *Comput. Sci. Rev.* **2023**, *49*, 100576. [CrossRef]
5. Patwardhan, N.; Marrone, S.; Sansone, C. Transformers in the real world: A survey on nlp applications. *Information* **2023**, *14*, 242. [CrossRef]
6. Rahali, A.; Akhloufi, M.A. End-to-end transformer-based models in textual-based NLP. *AI* **2023**, *4*, 54–110. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
8. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* **2023**, *241*, 122666. [CrossRef]
9. Sun, J.; Han, P.; Cheng, Z.; Wu, E.; Wang, W. Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis. *IEEE Access* **2020**, *8*, 211152–211163. [CrossRef]
10. Abas, A.R.; El-Henawy, I.; Mohamed, H.; Abdellatif, A. Deep Learning Model for Fine-Grained Aspect-Based Opinion Mining. *IEEE Access* **2020**, *8*, 128845–128855. [CrossRef]
11. Pereg, O.; Korat, D.; Wasserblat, M. Syntactically Aware Cross-Domain Aspect and Opinion Terms Extraction. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; International Committee on Computational Linguistics: New York, NY, USA, 2020; pp. 1772–1777. [CrossRef]
12. Kumar, A.; Veerubhotla, A.S.; Narapareddy, V.T.; Aruru, V.; Neti, L.B.M.; Malapati, A. Aspect Term Extraction for Opinion Mining Using a Hierarchical Self-Attention Network. *Neurocomputing* **2021**, *465*, 195–204. [CrossRef]
13. Dos Santos, B.N.; Marcacini, R.M.; Rezende, S.O. Multi-Domain Aspect Extraction Using Bidirectional Encoder Representations From Transformers. *IEEE Access* **2021**, *9*, 91604–91613. [CrossRef]

14. Li, R.; Chen, H.; Feng, F.; Ma, Z.; Wang, X.; Hovy, E. Dual Graph Convolutional Networks for Aspect-Based Sentiment Analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; Volume 1: Long Papers, pp. 6319–6329. [CrossRef]
15. Chen, C.; Teng, Z.; Wang, Z.; Zhang, Y. Discrete Opinion Tree Induction for Aspect-Based Sentiment Analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Chen, C., Teng, Z., Wang, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; Volume 1: Long Papers, pp. 2051–2064. [CrossRef]
16. Zhang, Z.; Ma, Z.; Cai, S.; Chen, J.; Xue, Y. Knowledge-Enhanced Dual-Channel GCN for Aspect-Based Sentiment Analysis. *Mathematics* **2022**, *10*, 4273. [CrossRef]
17. Mewada, A.; Dewang, R.K. SA-ASBA: A Hybrid Model for Aspect-Based Sentiment Analysis Using Synthetic Attention in Pre-Trained Language BERT Model with Extreme Gradient Boosting. *J. Supercomput.* **2023**, *79*, 5516–5551. [CrossRef]
18. Wu, F.; Li, X. Local Dependency-Enhanced Graph Convolutional Network for Aspect-Based Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 9669. [CrossRef]
19. Zhao, Q.; Yang, F.; An, D.; Lian, J. Modeling Structured Dependency Tree with Graph Convolutional Networks for Aspect-Level Sentiment Classification. *Sensors* **2024**, *24*, 418. [CrossRef]
20. Wang, P.; Tao, L.; Tang, M.; Wang, L.; Xu, Y.; Zhao, M. Incorporating Syntax and Semantics with Dual Graph Neural Networks for Aspect-Level Sentiment Analysis. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108101. [CrossRef]
21. Jiang, B. Heuristic-Enhanced Candidates Selection Strategy for GPTs Tackle Few-Shot Aspect-Based Sentiment Analysis. *arXiv* **2024**, arXiv:2404.06063. [CrossRef]
22. Verma, S.; Kumar, A.; Sharan, A. IAN-BERT: Combining Post-Trained BERT with Interactive Attention Network for Aspect-Based Sentiment Analysis. *SN Comput. Sci.* **2023**, *4*, 756. [CrossRef]
23. Wang, Z.; Xie, Q.; Feng, Y.; Ding, Z.; Yang, Z.; Xia, R. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. *arXiv* **2024**, arXiv:2304.04339. [CrossRef]
24. Kheiri, K.; Karimi, H. SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and Its Departure from Current Machine Learning. *arXiv* **2023**, arXiv:2307.10234. [CrossRef]
25. Zhang, H.; Cheah, Y.N.; Alyasiri, O.M.; An, J. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: A comprehensive survey. *Artif. Intell. Rev.* **2024**, *57*, 17. [CrossRef]
26. Jiang, Q.; Chen, L.; Xu, R.; Ao, X.; Yang, M. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6280–6285.
27. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 27–35. [CrossRef]
28. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Pavlopoulos, J.; Manandhar, S. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 486–495. [CrossRef]
29. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 19–30. [CrossRef]
30. Cai, H.; Xia, R.; Yu, J. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, 1–6 August 2021; Zong, C., Xia, F., Li, W., Naviglipp, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; Volume 1: Long Papers, pp. 340–350.
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [CrossRef]
32. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108. [CrossRef]
33. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv* **2019**, arXiv:1909.11942. [CrossRef]
34. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
35. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
36. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016. [CrossRef]

37. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17), Long Beach, CA, USA, 4–9 December 2017.
38. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV '17), Venice, Italy, 22–29 October 2017. [CrossRef]
39. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML '17), Sydney, Australia, 6–11 August 2017. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.