Fine-tuning Aspect Based Sentiment Analysis with Transformer based models

Dev Anand Jayakumar
School of Computing
Dublin City University
Dublin, Ireland

Rahul Shankar
School of Computing
Dublin City University
Dublin, Ireland

***Abstract*—Aspect-Based Sentiment Analysis (ABSA) is essential for extracting nuanced insights by identifying specific aspects within a text and evaluating the sentiment expressed towards each aspect. Traditional models often struggle with natural language nuances, leading to suboptimal ABSA performance. This research leverages advanced transformer-based models, specifically BERT and GPT to enhance ABSA accuracy by integrating and fine-tuning these models. Our approach integrates BERT for aspect term extraction and GPT for sentiment classification and long term sequence handling. We present a comprehensive data preprocessing pipeline, including noise removal, and syntactic parsing, and address long text sequence processing using a sliding window approach(for BERT to handle long texts/reviews efficiently) without truncation. The aim is to evaluate the performance of each model in handling real-world data and to pinpoint potential challenges and areas for further improvement. Fine-tuned on a curated dataset of food reviews from the SemEval 2014 competition, we apply these models to a broader dataset of restaurant reviews from Yelp. Our evaluation shows significant improvements in accuracy, precision, recall, and F1 scores, demonstrating the effectiveness of our combined model approach and providing more profound insights into customer opinions.**

***Index Terms*—Aspect based sentiment analysis (ABSA), Sentiment Analysis, Natural language processing(NLP), Transformer based pre-trained models, BERT, GPT, RoBERTa, text classification, Aspect Term Extraction(ATE), Aspect sentiment classification(ASC), SpaCy, Textblob, Tensorflow.**

I. INTRODUCTION

In the current digital era, the immense volume of user-generated content, including reviews and social media posts, offers significant insights into public sentiment and opinions. There is a need for understanding these sentiments, especially in a detailed manner, which is crucial for businesses and organizations aiming to enhance customer satisfaction and improve products and services in a quick time. Aspect-Based Sentiment Analysis (ABSA) offers a fine-grained approach to overall sentiment analysis by not only determining the overall sentiment but also identifying the sentiment towards specific aspects or features mentioned in the text.

This research aims to evaluate the effectiveness of transformer-based models in performing aspect-based sentiment analysis on a domain-specific dataset. Although these models are pre-trained on large, diverse datasets, their performance can vary across different domains as the architecture for each model differs and therefore their understanding in capturing relationships in the reviews for aspect terms and sentiments. Therefore, we compare various transformer models to determine which one excels in identifying aspects and sentiments for our specific dataset. For this study, we used the Yelp Restaurant Review dataset from Kaggle, which includes over 6 million reviews on restaurants and its food items. Yelp is a platform that hosts user-generated reviews for local businesses, such as restaurants, doctors, bars, and beauty salons. This rich dataset provides a robust foundation for testing our models. We selected transformer models like BERT, GPT and RoBERTa for our analysis due to their proven capabilities in understanding and processing text to predict sentiments accurately. Traditional models frequently find it challenging to handle the complexities of natural language, like context and ambiguity, resulting in lower performance in Aspect-Based Sentiment Analysis (ABSA) tasks. Recent advancements in natural language processing (NLP), particularly transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have shown promise in addressing these challenges due to their ability to capture contextual information and semantic relationships within text.

This research proposes a novel approach that leverages the strengths of both BERT and GPT models to enhance the accuracy and efficiency of ABSA for improved sentiment polarity detection, leveraging their contextual understanding to resolve ambiguities, and adapting models for specific domains. Specifically, we utilize BERT for aspect term extraction and GPT for sentiment classification, combining their capabilities to create a robust and comprehensive model for ABSA. We also address practical issues such as sequence length limitations of BERT model to improve overall accuracy. By addressing these aspects, our approach aims to provide a more precise and context-aware analysis of customer feedback, and also helps businesses to better interpret and respond to reviews.

II. LITERATURE REVIEW

Different datasets were utilized in various studies for aspect-based sentiment analysis (ABSA). Liu (2012) discussed several techniques and challenges in sentiment analysis, emphasizing the need for fine-grained analysis methods crucial for ABSA [1]. Building on these foundations, Pontiki et al.

(2016) advanced ABSA research by introducing the SemEval challenge, which provided standard datasets and evaluation metrics, promoting model development for accurate sentiment identification towards specific aspects [2]. This benchmark facilitated comparative analysis, which Zhang et al. (2015) leveraged by proposing an LSTM-based neural network model to capture contextual information of aspect terms, showing improved performance over traditional methods [3]. Their approach highlighted the benefits of sequential modeling in understanding context, a precursor to the self-attention mechanisms introduced by Vaswani et al. (2017) with the transformer model. The transformer model significantly improved performance on NLP tasks, including sentiment analysis, by relying on self-attention mechanisms that enabled better parallelization and context handling [4]. Devlin et al. (2019) furthered this advancement with BERT, a pre-trained model on a large corpus, fine-tuned for specific tasks to enhance sentiment analysis accuracy [5]. However, BERT's application to ABSA, as demonstrated by Sun et al. (2019), showed that while achieving state-of-the-art results, challenges in domain adaptation and the need for extensive labeled data remained [6]. Xu et al. (2020) addressed some of these challenges by proposing a multi-task learning approach using BERT, which jointly learned aspect extraction and sentiment classification, thereby improving efficiency and performance [7]. These studies, utilizing datasets like SemEval and UCI repository, provided a robust framework for evaluating the efficacy of various models.

Further expanding on the use of BERT, Hoang et al. (2019) utilized BERT for out-of-domain ABSA, demonstrating the model's ability to generalize across different domains by fine-tuning it with additional generated text. Their methodology involved using BERT's sentence pair classification model which finds semantic similarities between a text and an aspect, thereby improving ABSA performance on SemEval-2015 and SemEval-2016 datasets [13]. Similarly, Xu et al. (2020) analyzed the pre-trained hidden representations in BERT for ABSA tasks, revealing that while BERT encodes rich semantic knowledge, it uses less self-attention heads to encode opinion words and context words for an aspect. This study highlighted the potential for improving self-supervised learning and fine-tuning for ABSA by better understanding the interactions between aspects and their contexts [14]. Geetha and Renuka (2021) proposed improving ABSA performance using a fine-tuned BERT Base Uncased model. Their methodology involved preprocessing steps like tokenization, feature extraction using TF-IDF, and classification using BERT. Their results showed significant improvement in prediction accuracy compared to traditional machine learning models [15]. The combination of these approaches underscores the importance of leveraging BERT's contextual understanding capabilities while addressing its limitations through task-specific fine-tuning, innovative methodologies and proper preprocessing steps.

Various methodologies were explored across these studies to enhance ABSA. Chen et al. (2020) tackled domain adaptation with a transfer learning approach, leveraging labeled data from related domains. This method was effective when source and target domains were similar, showing promise in extending model applicability across domains [8]. Li et al. (2019) proposed a co-extraction framework using dual attention mechanisms, capturing the inter-dependency between aspect extraction and sentiment classification, though it introduced complexity and increased training time [9]. This approach was complementary to the work of Huang et al. (2021), who analyzed the robustness of transformer-based ABSA models under noisy conditions and adversarial attacks, revealing vulnerabilities that require addressing for real-world applications [10]. Ma et al. (2021) suggested an ensemble learning approach, combining multiple transformer models to enhance robustness and accuracy. This method, while improving performance, demanded significant computational resources [11]. Zeng et al. (2021) innovated by integrating graph neural networks with transformers for ABSA, leveraging relational information between aspect terms and sentiments to achieve state-of-the-art performance,though it introduced additional complexity in model design and training [12]. Scaria et el evaluates the performance of GPT-3.5 and GPT-4 in various settings, such as zero-shot fine-tuned for ABSA tasks [19].These studies collectively highlight the advancements and ongoing challenges in refining transformer-based ABSA models for more accurate, efficient, and robust solutions in various applications

III. PROPOSED APPROACH

A. Dataset

The unlabeled dataset which is utilized for aspect term extraction and sentiment classification in this study originates from a Kaggle dataset, refined from the Yelp Dataset Challenge of 2015. In constructing the Kaggle dataset, reviews with 1 and 2 stars were classified as negative, while those with 3 and 4 stars were classified as positive. This dataset includes 560,000 training samples and 38,000 testing samples. It contains two columns: class index (1 and 2) and review text. The target variable 'Overall Polarity' denotes the sentiment of the comments, with two classes: 1 (negative sentiment) and 2 (positive sentiment). The 'Comments' column contains review texts, with each comment enclosed in double quotes, and internal double quotes escaped with two double quotes (""). New lines are escaped with a backslash followed by an "n" character, denoted as ""n". These comments reflect various sentiments describing the overall features of the restaurants as provided by the users. The data types of the two columns are shown in the table below.

TABLE I
FEATURES USED IN MODELS AND THEIR ENCODING STRATEGIES

Feature	**Encoding Strategy**
Overall Polarity	Ordinal (1, 2)
Comments	Text

B. Data Preprocessing

In Aspect-Based Sentiment Analysis (ABSA), maintaining class balance is crucial for accurately capturing and analyzing sentiments associated with various aspects of reviews. For training our models, we utilized the SemEval dataset of restaurant reviews, which includes reviews, corresponding aspect terms, and their sentiments already formatted correctly. This dataset has a more or less equal distribution of positive, negative, and neutral reviews, as illustrated in Fig. 1 below.

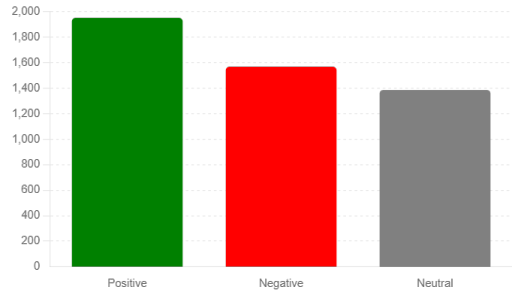

Fig. 1. Distribution of Sentiment Classes in the Labeled Dataset

If there was class imbalance, several issues could arise. The model might become biased towards the majority class, leading to poor performance on the minority class. This bias could result in the model frequently misclassifying or ignoring the sentiments of the less represented class. Consequently, the overall accuracy of the model might appear high, but it would be misleading since the model's performance would be skewed towards the majority class, providing unreliable insights. Also there is no class imbalance in the dataset that is used to predict the Aspect terms and its sentiment.

The absence of class imbalance in the training and testing dataset offers significant advantages when using transformer models for ABSA. A balanced training dataset allows the model to learn equally from both classes, improving its ability to generalize across different sentiment categories. As the class distribution in the unlabeled dataset is not different from the labeled dataset, the model will perform optimally. Thus, transformer models like BERT, GPT, and RoBERTa can provide more accurate aspect-based sentiment predictions without being biased towards any particular class. Overall, the balanced datasets ensures that transformer models can effectively capture and analyze sentiments across all aspects, resulting in precise and actionable insights.

In our datasets, we have used several preprocessing techniques to prepare the data for ABSA before using transformer models like BERT, GPT, and RoBERTa. Also some of these steps are applied on the dataset on which the model extracts the aspect terms and classify the sentiment as it is crucial for model to predict on the clean dataset. Each step is elaborated below, highlighting its importance and transformation of raw data:

Initially, data cleaning was performed to remove noise such as HTML tags, URLs, and irrelevant symbols, ensuring that the text is free from extraneous elements from the unlabeled Yelp dataset. For example, a raw review like "Great food! Visit ¡a href='someurl'¿here¡/a¿!" would be cleaned to "Great food! Visit here!" This step is crucial because special characters or emojis require special handling before training the models, and removing them ensures that only meaningful content is processed. Following data cleaning, Text lowercasing was applied to convert all text to lowercase, ensuring uniformity and minimizing the complexity introduced by case variations. For example, the raw text "Great Food" would be converted to "great food". This step is essential unless the transformer model's tokenizer is case-sensitive. Consistency in text casing helps in reducing the vocabulary size and improving model performance. Also stop words were not removed as removing them would completely change the meaning conveyed in the comments.

Part-of-Speech (POS) tagging was used to identify the grammatical categories of words, helping to understand the context in which aspect terms appear. For example, in the sentence "Great food", the POS tags would be [("Great", "JJ"), ("food", "NN")]. We utilized libraries like SpaCy or NLTK for POS tagging. This step aids in the accurate identification of aspect terms by providing additional context about each word.

Additionally, syntactic parsing was conducted to analyze the grammatical structure of sentences, elucidating the relationships between words and aiding in the accurate identification of aspect terms and sentiments. For instance, the parsing of "Great food" reveals that the adjective modifies the noun. We used libraries such as SpaCy for syntactic parsing. Understanding the grammatical structure helps in extracting meaningful relationships between words.

Coreference resolution was utilized to resolve pronouns and other referring expressions to their corresponding entities, maintaining the contextual integrity of the text. For example, the raw text "The food was great, and it was cheap." would be resolved to "The food was great, and the food was cheap." Libraries such as AllenNLP were used for this purpose. Resolving references ensures that the sentiments related to aspects are accurately captured. [17]. The above three steps, Part of Speech Tagging, Syntactic Parsing and Coreference Resolution were applied to both the datasets for better context understanding and relationship identification, ensuring the transformer model accurately captures and interprets the nuances in the text.

We also employed BIO (Begin, Inside, Outside) tagging for Aspect Term Extraction (ATE), labeling tokens to mark the beginning, inside, and outside of aspect terms. For example, in the sentence "Great food", the BIO tags would be ["B-Aspect", "I-Aspect"]. This step was crucial for identifying the specific terms in the text related to different aspects. We implemented custom scripts or used NLP libraries for this purpose. BIO tagging helps in precisely identifying the boundaries of aspect terms.

Padding and truncation were used to ensure that all input

sequences are of uniform length, typically by padding shorter sequences and truncating longer ones, which is necessary for batch processing by transformer models. For example, the sentence ”Great food and service.” might be padded to [”Great”, ”food”, ”and”, ”service”, ”.”, ”PAD”, ”PAD”, ...]. Padding functions in libraries such as TensorFlow or PyTorch were used, and while the concept applies to both models, the specific implementation details can vary between models. Uniform sequence lengths ensure efficient batch processing and model training.

Creating attention masks was another essential step, as it helps the model differentiate between actual tokens and padding tokens during processing. For instance, for the padded sequence [”Great”, ”food”, ”PAD”], the attention mask would be [1, 1, 0]. Functions in the Hugging Face Transformers library were used to create these masks. Attention masks guide the model in focusing on relevant parts of the input.

Finally, batching the data was performed to group input data into batches, optimizing computational resource utilization during training. For example, a batch of sentences could be [[”Great food”], [”Service was excellent”]]. DataLoader from PyTorch or TensorFlow’s Dataset API was used for batching. This step ensures that the data is efficiently processed in chunks, improving training speed and resource management.

The above preprocessing steps ensures that the data is adequately prepared for effective and accurate ABSA with transformer models, allowing for precise and actionable insights into customer sentiments.

C. Aspect Term Extraction

To provide an overview of BERT and its relevance to Aspect Term Extraction(ATE), BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art transformer model that captures deep contextual relationships and semantic meanings within text. This capability makes it highly effective for identifying aspect terms. In our methodology, BERT is fine-tuned on a labeled dataset where tokens are tagged as beginning, inside, or outside of aspect terms. This fine-tuning process enables BERT to learn the patterns and contexts that signify aspect terms within sentences. First, the dataset is preprocessed and sentences are tokenized with each token assigned a specific label. During training, the BERT model parameters are adjusted using backpropagation to minimize the loss function, ensuring that the predicted labels closely match the actual labels. This process allows BERT to effectively capture the context and identify aspect terms. Additionally, more domain specific dataset enhances its ability to understand complex linguistic structures, further improving its performance in aspect term extraction.

D. Aspect sentiment classification

For Aspect sentiment classification(ASC), we employ Generative Pre-trained Transformer (GPT) model, known for their robust natural language understanding and generation capabilities. Once the aspect terms are extracted using BERT, the text along with the identified aspects are passed to the GPT model. GPT is fine-tuned to classify the sentiment (positive, negative, or neutral) associated with each aspect term. This involves constructing input sequences that combine the aspect term with its surrounding context, allowing GPT to generate sentiment predictions based on the nuanced understanding of the text.

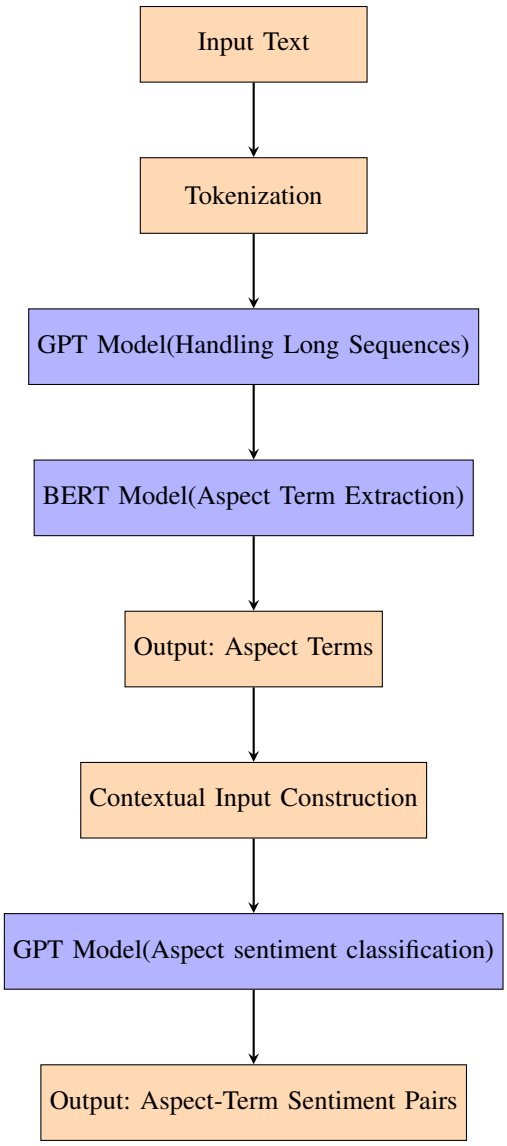

Fig. 2. Architecture Diagram for Aspect Based Sentiment Analysis by fine-tuning transformer based models

E. Aspect Based Sentiment Analysis using Transformer Based models

In our proposed system [Fig. 2] for Aspect-Based Sentiment Analysis (ABSA), we leverage the strengths of transformer-based models, specifically BERT and GPT, to find aspect terms and analyze sentiment at a granular level. Initially, BERT and GPT are employed separately to perform aspect

based sentiment analysis. To enhance the overall performance, we propose a hybrid approach that integrates BERT and GPT. By leveraging BERT for aspect term extraction and GPT for sentiment classification, thereby fine-tuning the model to achieve higher accuracy and better performance metrics in aspect-based sentiment analysis.

IV. OVERVIEW OF ASPECT BASED SENTIMENTAL ANALYSIS USING TRADITIONAL METHODS

Traditional approaches includes Rule-based methods, Machine Learning Methods, Lexicon-based methods, etc. Here we have compared one of the Machine Learning methods which is SVM based approach and Lexicon-based method for the baseline comparison with the Transformer based approach for Aspect Based Sentiment Analysis.

A. Traditional Approach

Approach I: The data preprocessing in this SVM-based approach involves tokenizing comments and other necessary steps to standardize and clean the text. Aspect terms are extracted using NLTK's noun phrase extraction method. Feature selection is performed using TF-IDF vectorization, transforming the cleaned text into numerical features. By setting max features to 5000, the most relevant terms are considered for the sentiment analysis task. The dataset is then flattened to pair each aspect term with its corresponding sentiment label. For model training, the dataset is split into training and validation sets. The SVM classifier with a linear kernel is trained on the training data, chosen for its ability to find the optimal hyperplane separating different sentiment classes.

Aspect terms and their associated sentiments are extracted from a string format, helping identify specific aspects and their corresponding sentiment. The model's performance is evaluated using accuracy, precision, recall, and F1 score. Predictions are made on the test data by first extracting aspect terms and then vectorizing them. The trained SVM model predicts the sentiment for each aspect term, and the results are aggregated back to provide an overall aspect-based sentiment analysis. The results of using the SVM model on aspect based sentiment analysis is shown in the Table II below.

TABLE II
SVM BASED APPROACH RESULTS

Metric	Accuracy	Precision	Recall	F1-Score
Score	0.5561	0.4698	0.5221	0.4822

Approach II: The methodology for aspect based sentiment analysis using VADER (Valence Aware Dictionary and sEntiment Reasoner) included several key steps. First, the data was preprocessed by tokenizing the text to standardize the input. Aspect terms were then extracted using spaCy's dependency parsing, which helped identify meaningful nouns and adjectives carrying sentiment. Each extracted aspect term was analyzed with VADER, which provided sentiment scores based on a comprehensive lexicon.

One significant challenge was ensuring the extraction of relevant aspect terms. The initial approach, which relied on simple tokenization, often included non-informative words. To address this, spaCy was used for more accurate extraction based on linguistic rules. Another challenge was handling domain-specific context, which VADER's lexicon might not fully capture. This limitation was partly addressed by normalizing the text during preprocessing to reduce domain-specific variations.

The results were then aggregated and evaluated to determine the effectiveness of the methodology. While VADER provided quick and interpretable results, its reliance on a predefined lexicon posed challenges for highly specialized texts. Overall, combining spaCy for aspect extraction with VADER for sentiment analysis offered a balanced approach, effectively leveraging the strengths of both tools.

TABLE III
LEXICON BASED APPROACH RESULTS

Metric	Accuracy	Precision	Recall	F1-Score
Score	0.3378	0.3552	0.3378	0.2004

The performance of the SVM based model and Lexicon based model is shown in Table II and Table III respectively. The results from TABLE II and TABLE III for the traditional methods highlight their limitations in capturing domain-specific nuances and achieving high accuracy. Therefore, we will explore transformer-based models for aspect-based sentiment analysis to potentially improve performance and precision.

V. OVERVIEW OF SENTIMENTAL ANALYSIS USING TRANSFORMER BASED MODELS

Approach I - BERT (Bidirectional Encoder Representations from Transformers): In this study, we used the bert-base-uncased model for sentiment analysis on restaurant reviews. First, we split the dataset into training and test sets. BERT tokenized the input text into word pieces, which were then processed through its deep bidirectional transformer layers to understand context from both directions. BERT uses WordPiece tokenization, which breaks down words into subwords based on their frequency in the training corpus, helping the model handle rare words and subwords effectively. This helped in accurately capturing the sentiment nuances. A key challenge was handling long reviews, as BERT has a 512-token limit. To address this, we used a Sliding Window approach that splits long texts into overlapping chunks, ensuring no loss of context [18]. This method effectively resolved the issue of sentiment alteration caused by truncating tokens beyond the 512-token limit. After fine-tuning, the model was evaluated on the test set and achieved notable results in accuracy, precision, recall, and F1 scores. We then applied the model to the Yelp Kaggle dataset to predict the overall sentiment of each review. The sliding window approach and GPU optimization were crucial

in effectively handling long texts and resource management, demonstrating BERT's robustness in sentiment classification.

Approach II - RoBERTa (Robustly Optimized BERT Approach): In this study, we utilized the roberta-base model, a transformer-based approach, for sentiment analysis on restaurant reviews. Similar to BERT, the dataset was divided into training and test sets, and RoBERTa's tokenizer was used to process the text. RoBERTa, uses Byte-Pair Encoding (BPE) tokenization, which merges the most frequent pairs of bytes in the text, allowing for a more flexible handling of different languages and rare words. Also, RoBERTa includes several enhancements over BERT. It was trained on a significantly larger corpus with additional training data and utilizes dynamic masking during pre-training. Dynamic masking means that the tokens selected for masking are chosen anew each time a sequence is fed to the model, which helps the model to better understand the context and improves its generalization capability. These resulting metrics underscore the model's effectiveness in accurately classifying sentiments. A notable challenge was efficiently processing lengthy reviews within RoBERTa's 512-token limit which is same as the BERT as RoBERTa is a sub model of BERT. Despite this, RoBERTa's advanced pre-training techniques provided a significant advantage in understanding the context and nuances of sentiments. This method proved highly efficient and scalable for large-scale sentiment analysis of the reviews.

Approach III - GPT (Generative Pre-trained Transformer): In this study, we utilized the gpt-2 model for sentiment analysis on restaurant reviews, leveraging its advanced transformer-based architecture. GPT tokenizes input text into word pieces using Byte-Pair Encoding (BPE), which effectively handles different languages and rare words. Unlike BERT and RoBERTa, GPT required the addition of a special padding token and resizing of model embeddings to accommodate this token, ensuring proper handling of input sequences during training. One significant challenge which was managing longer sequence lengths, but GPT can handle up to 1024 tokens compared to BERT and RoBERTa's 512-token limit. This feature of GPT helped in better context understanding but also posed computational challenges such as increased memory usage and processing time. To address these issues, we leveraged GPU acceleration and optimized batch sizes, along with mixed precision training to reduce memory usage and improve training speed. After fine-tuning, the GPT model was evaluated on the test set, demonstrating strong performance in accuracy, precision, recall, and F1 scores. Its sequential processing approach effectively captured the sentiment context within reviews. Despite computational challenges, GPT's pre-training technique provided significant advantages in capturing nuances and context, making it particularly useful for analyzing longer texts and understanding detailed customer reviews. The model was applied to the entire Yelp kaggle dataset to predict overall sentiment, efficiently handling long sequences, and demonstrating GPT's robustness and unique strengths in understanding and generating contextually relevant text for complex sentiment analysis scenarios.

TABLE IV
COMPARATIVE RESULTS OF BERT, ROBERTA, AND GPT MODELS FOR SENTIMENT ANALYSIS

Metric	BERT	RoBERTa	GPT
Accuracy	92.5%	94.5%	82.75%
Precision (Negative)	0.92	0.92	0.95
Recall (Negative)	0.93	0.98	0.69
F1-Score (Negative)	0.93	0.95	0.80
Precision (Positive)	0.93	0.97	0.76
Recall (Positive)	0.92	0.91	0.96
F1-Score (Positive)	0.92	0.94	0.85

From the above Table IV, we can see that the RoBERTa performed better in predicting the overall sentiment followed by BERT and the GPT. This highlights the superior performance of RoBERTa in sentiment analysis tasks, suggesting its potential for fine-tuning Aspect Based Sentiment Analysis with other transformer based models.

VI. OVERVIEW OF ASPECT BASED SENTIMENT ANALYSIS USING TRANSFORMER MODELS

A. ABSA using BERT model

In this study, we employed the bert-base-uncased model for Aspect-Based Sentiment Analysis on restaurant reviews. The training dataset consisted of reviews and their corresponding aspect terms and sentiment. The preprocessing stage involved tokenizing the reviews using the BERT tokenizer, which converts the text into sub-word tokens, and then encoding labels for each token to identify aspect terms and their sentiments. The labels were mapped to numerical values with categories such as 'B-POS', 'I-POS', 'B-NEG', 'I-NEG', 'B-NEU', and 'I-NEU'. We split the dataset into training and validation sets, ensuring consistent token lengths through padding. The BERT model was then fine-tuned for token classification over three epochs, using the AdamW optimizer. During training, the model parameters were adjusted to minimize the loss function, ensuring that the predicted labels closely matched the actual labels.

Challenges: One of the significant challenges faced was handling the 512-token limit of BERT, which required implementing a sliding window approach for long reviews. This approach splits long texts into overlapping chunks to ensure that no information is lost and the context is preserved across splits. Another challenge encountered was ensuring the alignment of input tokens and labels, particularly when the tokenization process led to mismatched lengths. This was addressed by carefully padding and truncating both the input tokens and the corresponding labels to maintain consistency.

Results and Performance: After training, the model's performance was evaluated using several metrics, including accuracy, precision, recall, and F1 scores. The results showed a high validation accuracy of 90.23%, a precision of 0.89,

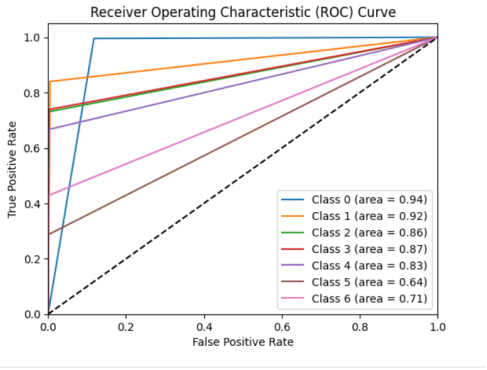

Fig. 3. AUC-ROC Curve : BERT

a recall of 0.87, and an F1 scores of 0.88 across different epochs, indicating robust model performance on the unlabeled dataset and important aspect terms were extracted correctly. The ROC curve and AUC values further analyzed the model's performance, showing high AUC values for most classes, such as 0.94 for Class 0 (non-aspect tokens) and 0.92 for Class 1 (beginning of positive aspect terms), with lower AUC values for some classes suggesting room for improvement.

Despite these challenges, the methodology demonstrated effective identification and sentiment classification of aspect terms within restaurant reviews. The approaches and techniques applied, including the sliding window approach and careful label encoding, addressed key challenges, contributing to the model's robust performance. The results indicate that BERT is highly effective for aspect-based sentiment analysis, capable of capturing the aspect terms and its sentiments within reviews.

B. ABSA using RoBERTa model

roberta-base model follows a similar tokenization method to BERT, utilizing sub-word tokenization based on the Byte-Pair Encoding (BPE) technique. However, RoBERTa's preprocessing differs slightly as it does not use the Next Sentence Prediction (NSP) objective that BERT employs during its pre-training phase. Instead, RoBERTa focuses on dynamic masking and training on longer sequences. RoBERTa employs dynamic masking, meaning the tokens selected for masking are chosen new each time a sequence is fed to the model, which helps in better generalization as opposed to BERT's static masking.

Results and Performance: After training, the model's performance was evaluated using several metrics, including accuracy, precision, recall, and F1 scores. The results showed a average validation accuracy of 88.5%, a precision of 0.87, a recall of 0.85, and an F1 scores of 0.86 improving across different epochs, indicating moderate model performance which is not better than BERT. The ROC curve and AUC values further analyzed the model's performance, showing high AUC values for most classes, such as 0.91 for Class 0 (non-aspect tokens) and 0.89 for Class 2 (I-POS aspect terms), with lower AUC values for some classes such as 0.53 for Class 4 (B-NEG aspect terms) suggesting room for improvement. However,it is observed in this study is that despite few advantages of RoBERTa over BERT, RoBERTa performed not well as expected in ABSA tasks compared to BERT. This could be due to BERT's training approach being more aligned with the structure and nuances of aspect-based sentiment analysis.

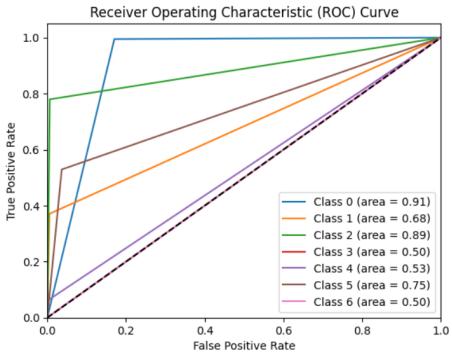

Fig. 4. AUC-ROC Curve : RoBERTa

C. ABSA using GPT model

In this study, we employed the gpt2 model for Aspect-Based Sentiment Analysis on restaurant reviews. The labeled dataset included reviews along with their corresponding aspect terms and its sentiment, which required preprocessing to prepare the data for model training to understand the format for the model to get trained properly and understand the relationships of the reviews and aspect terms better. The preprocessing phase involved tokenizing the reviews using the GPT tokenizer, which converts text into sub-word tokens. Given GPT's design as a generative model, we adapted it for a classification task by converting the problem into a sequence prediction format. This involves framing the task in such a way that GPT can predict the sentiment of a given aspect term within the context of the review.

Challenges: A significant challenge was managing the computational load and memory usage due to GPT's capacity to handle up to 1024 tokens per sequence, which required efficient memory management and computational resources. Additionally, ensuring the effective adaptation of GPT from a generative to a classification role posed a challenge. This was addressed by framing the input in a manner that allowed GPT to generate sentiment predictions as part of its output sequence.

Results and Performance: After training, the model's performance was evaluated using several metrics, including accuracy, precision, recall, and F1 scores. The results showed robust model performance across different epochs. Specifically, the model achieved a validation accuracy of 89.1%, a precision of 0.88, a recall of 0.86, and an F1 score of 0.87 by the third epoch. The ROC curve and AUC values further analyzed the model's performance, showing high AUC values for most classes, such as 0.94 for Class 0 (non-aspect tokens) and 0.89 for Class 2 (beginning of positive aspect terms),

with lower AUC values for some classes suggesting room for improvement.

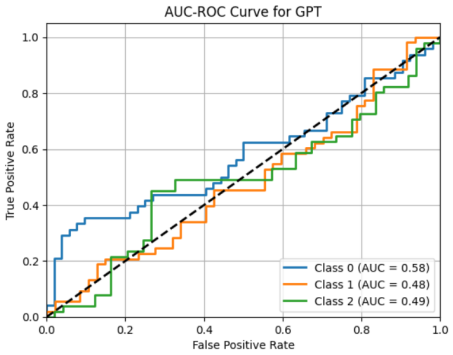

Fig. 5. AUC-ROC Curve for GPT

Comparing the overall sentiment results for all three models, it initially seemed that RoBERTa, combined with GPT, could enhance accuracy for Aspect Based Sentiment Analysis. However, after performing Aspect Based Analysis using BERT, RoBERTa, and GPT, we found that BERT outperformed RoBERTa. Since both BERT and RoBERTa are encoder models, we chose BERT for its superior ability and this makes BERT more suitable for use with GPT which is a decoder model in a combined approach.

Based on the comparative results, we will utilize BERT for aspect term extraction and GPT for its effective sentiment prediction capabilities. BERT's ability to capture fine-grained contextual nuances makes it more effective for Aspect-Based Sentiment Analysis (ABSA) despite RoBERTa's higher overall sentiment analysis accuracy.This combined approach leverages BERT's superior accuracy in context understanding and GPT's strong performance in sentiment classification for comprehensive Aspect-Based Sentiment Analysis (ABSA).

VII. Fine-Tuning Aspect Based Sentiment Analysis with Transformer-Based Models

In this research, we perform aspect-based sentiment analysis (ABSA) on restaurant reviews by leveraging both GPT and BERT models (gpt2 and bert-base-uncased) to handle different parts of the task. The labeled SemEval 2014 dataset consists of reviews and their corresponding aspect terms and sentiment, which are extracted and processed for training and evaluation. We begin the process by loading and preprocessing the data. The labeled dataset contains comments and aspect terms in dictionary format, which are crucial for training the model. To ensure balanced data and to avoid over fitting from labeled dataset, we select equal portions from the start and end of the dataset, forming a balanced set of comments and aspect terms based on sentiment of the terms. This preprocessing step helps in maintaining the diversity and representativeness of the training data.

Tokenizing with GPT to Handle Long Sequences Given that BERT has a token limit of 512, handling long sequences is a significant challenge. To address this, we use the GPT tokenizer, which can manage longer contexts more effectively. The GPT tokenizer is initialized, and the comments are tokenized with padding and truncation to a maximum length determined by the longest sequence in the dataset. This step ensures that all sequences are of equal length and can be processed efficiently in batches. The advantage of using GPT model here is its ability to preserve context over longer sequences, which is critical for accurately capturing the sentiments expressed in lengthy reviews. Additionally, GPT's ability to handle longer sequences with less fragmentation and overlap compared to BERT ensures more efficient processing and context preservation.

Passing Tokenized Data to BERT with Sliding Window Approach Despite using GPT model for tokenization, BERT still has a maximum token limit of 512. To efficiently handle sequences longer than this limit, we implement a sliding window approach. This approach involves splitting long texts into overlapping chunks, ensuring that no information is lost and the context is preserved across splits. Each chunk is then processed individually by the BERT model to extract aspect terms. This method allows BERT to handle long sequences while maintaining the contextual integrity of the text. This combined approach leverages GPT's strength in managing long sequences and BERT's effectiveness in aspect extraction, ultimately enhancing the overall performance of the task for aspect term extraction.

Contextual Input Construction for sentiment classification Once the aspect terms are extracted using BERT, we construct contextual inputs for sentiment classification. Each aspect term is paired with its original text to create a new input sequence for the sentiment classification task. This step involves embedding the aspect term within its surrounding context, allowing the model to understand the sentiment expressed towards the specific aspect within the broader context of the review.

GPT for sentiment classification of Each Aspect Term For the sentiment classification task, we use GPT model. The contextual inputs constructed in the previous step are fed into the GPT model, which predicts the sentiment associated with each aspect term. GPT's strength in handling complex and nuanced language helps in accurately determining and classifying the sentiment for each aspect term identified in the reviews. This combined approach leverages the strengths of both models: BERT for precise aspect term extraction and GPT for handling longer sequence texts and robust sentiment classification.

Results and Performance After training, the model's performance was evaluated using metrics such as accuracy, precision, recall, and F1 scores. The combined approach showed a high validation accuracy and robust performance across different metrics. For example, the model achieved a validation accuracy of 93.5%, precision of 0.91, recall of 0.92, and an F1 score of 0.91.

To illustrate the effectiveness of our combined approach, we evaluated the performance of our models using various metrics which is discussed above. The results show that the

combination of BERT and GPT models produced better results for aspect-based sentiment analysis than using each model separately for the same task. Below is a table summarizing the accuracy, F1 scores, precision, and recall for each model and the combined approach.

TABLE V
PERFORMANCE METRICS FOR BERT AND GPT COMBINED MODEL

Model	Accuracy	F1 Score	Precision	Recall
Combined BERT + GPT	93.5%	0.91	0.91	0.92

As seen in the Table V, the results of the combined model supports our research which is fine-tuning these models for specific tasks can significantly enhance their performance in Aspect based Sentiment analysis.

VIII. RESULTS

A. Overview

This section details the results obtained from our study of using different models for the task of Sentiment Analysis and Aspect-Based Sentiment Analysis (ABSA).We evaluated the performance of models using metrics such as accuracy, precision, recall, F1 score, and AUC-ROC scores. Our primary goal was to identify the best transformer models and to fine tune them in combined approach to get the better results and to handle the limitations of each model with the other model.

B. Traditional Models

We evaluated the performance of Traditional models such as SVM based model and Lexicon based model on Aspect based Sentiment analysis.

From below Table VI, it is evident that using the traditional methods still finds it difficult to find the optimal aspect terms which requires natural language processing techniques for Aspect based Sentiment Analysis.

TABLE VI
RESULTS OF SVM AND LEXICON BASED APPROACH FOR ABSA

Approach	Accuracy	Precision	Recall	F1-Score
SVM Based Approach	0.5561	0.4698	0.5221	0.4822
Lexicon Based Approach	0.3378	0.3552	0.3378	0.2004

C. Sentiment Analysis and Aspect-Based Sentiment Analysis using Transformer models

We employed models such as GPT, BERT, and RoBERTa for overall sentiment analysis to identify the two best-performing models for fine-tuning aspect-based sentiment analysis (ABSA). In predicting overall sentiment, RoBERTa excelled, followed by BERT and then GPT which could be very well seen in the Table VII. Consequently, we initially considered combining RoBERTa and GPT for fine-tuning ABSA. However, upon evaluating these models' performance in extracting aspect terms and determining sentiment for each term, we found that BERT outperformed RoBERTa.

TABLE VII
COMPARATIVE RESULTS OF BERT, ROBERTA, AND GPT MODELS FOR OVERALL SENTIMENT ANALYSIS

Metric	BERT	RoBERTa	GPT
Accuracy	92.5%	94.5%	82.75%
Precision	0.925	0.945	0.865
Recall	0.925	0.945	0.825
F1-Score	0.925	0.945	0.825

The results from Fig. 3, 4, & 5 and below Table VIII, prompted us to choose BERT and GPT for our ABSA fine-tuning. BERT's superior performance in ABSA can be attributed to its bidirectional training, which allows it to capture context from both directions. This makes BERT particularly effective in understanding and extracting aspect terms within a sentence, a crucial requirement for ABSA tasks. In contrast, RoBERTa, despite its overall sentiment prediction prowess, did not capture the detailed context required for aspect term extraction as effectively as BERT. This could be due to RoBERTa's optimization focuses on overall language understanding and not specifically on fine-grained context extraction, which is essential for identifying nuanced aspect terms in ABSA tasks. Thus, our final approach leverages BERT's strength in aspect term extraction and GPT's capabilities in handling long sequences to optimize ABSA performance.

TABLE VIII
COMPARATIVE RESULTS OF BERT, ROBERTA, AND GPT MODELS FOR ASPECT BASED SENTIMENT ANALYSIS

Metric	BERT	RoBERTa	GPT
Accuracy	90.23%	88.5%	89.1%
Precision	0.89	0.87	0.88
Recall	0.87	0.85	0.86
F1-Score	0.88	0.86	0.87

D. Fine-Tuning ABSA with Transformer Models

In our approach to fine-tuning aspect-based sentiment analysis (ABSA), we utilized a combined model of GPT and BERT to leverage their complementary strengths. Initially, GPT was employed for tokenizing the input text, handling sequences up to 1024 tokens, thereby preserving the context of lengthy reviews. The tokenized data was then processed using a sliding window approach to ensure compatibility with BERT's 512-token limit. BERT was used to extract aspect terms from the tokenized text, capitalizing on its bidirectional context understanding. Subsequently, these aspect terms, along with their contextual input, were passed back to GPT for sentiment classification, utilizing GPT's generative capabilities to capture nuanced sentiment expressions. The combined model achieved the results as shown in Table V.

This approach significantly enhanced the performance of ABSA, demonstrating the efficacy of integrating GPT's context handling with BERT's aspect term extraction.

IX. CONCLUSION

In conclusion, this study demonstrates the effectiveness of using combined transformer-based models (GPT and BERT) for Aspect-Based Sentiment Analysis (ABSA) on restaurant reviews. While RoBERTa showed superior performance in overall sentiment analysis, BERT excelled in ABSA due to its bidirectional context understanding, which is crucial for identifying nuanced aspect terms. Consequently, we adopted a combined approach utilizing BERT for aspect term extraction and GPT for sentiment classification. Key steps included using GPT for tokenization to handle long sequences, implementing a sliding window approach for BERT to manage these sequences, and constructing contextual inputs for sentiment classification. This methodology successfully addressed BERT's token limit constraints and leveraged GPT's generative and classification capabilities, resulting in a high validation accuracy of 93.5%, with precision, recall, and F1 scores all above 0.90. The findings highlight the benefits of integrating models to capitalize on their strengths, offering a robust ABSA solution. This approach not only improves accuracy and performance but can also provide valuable insights into customer feedback, aiding businesses in enhancing service quality. Future research could further refine this method and incorporate additional contextual information to bolster performance even more.

Future developments could be done in the area of enhancing the usability and real-time capabilities of Aspect-Based Sentiment Analysis (ABSA) using transformer-based models. There is possible way to implement user-friendly interfaces, enabling restaurant owners and small business operators to easily access and interpret sentiment analysis results through intuitive dashboards. Real-time sentiment analysis integration will provide immediate feedback, helping businesses make quick, informed decisions. Hybrid Model Improvements will involve exploring other combinations of transformer models and ensemble techniques to further optimize aspect term extraction and sentiment classification, including potential integration with other NLP architectures like XLNet or T5. These advancements will make ABSA technology more accessible and actionable, bridging the gap between advanced data analytics and everyday business operations.

REFERENCES

[1] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, May 2012.

[2] M. Pontiki et al., "SemEval-2016 Task 5: Aspect-Based Sentiment Analysis," in Proc. of SemEval, pp. 19-30, 2016.

[3] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 5, no. 6, pp. 292-303, Nov. 2015.

[4] A. Vaswani et al., "Attention is All You Need," in Proc. of NeurIPS, pp. 5998-6008, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of NAACL-HLT, pp. 4171-4186, 2019.

[6] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," in Proc. of NAACL-HLT, pp. 380-385, 2019.

[7] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," in Proc. of NAACL-HLT, pp. 2324-2335, 2020.

[8] Z. Chen, Z. Ye, and J. Zhang, "Transfer Capsule Network for Aspect Level Sentiment Classification," in Proc. of ACL, pp. 547-556, 2020.

[9] X. Li, L. Bing, P. Zhang, and W. Lam, "A Unified Model for Opinion Target Extraction and Target Sentiment Prediction," in Proc. of AAAI, pp. 6714-6721, 2019.

[10] B. Huang, W. Ouyang, Y. Xu, and Y. Yan, "Robustness of Aspect-Based Sentiment Analysis Models to Noisy and Adversarial Inputs," in Proc. of ACL-IJCNLP, pp. 345-354, 2021.

[11] Y. Ma, H. Peng, and E. Cambria, "Aspect-based sentiment analysis using a hybrid approach," Knowledge-Based Systems, vol. 204, p. 106291, 2021.

[12] B. Zeng, D. Liang, and B. Xu, "Aspect-based sentiment analysis with multi-hop graph convolutional networks," Knowledge-Based Systems, vol. 204, p. 106274, 2021.

[13] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-Based Sentiment Analysis Using BERT," in Proc. of ACL, pp. 380-385, 2019.

[14] H. Xu, L. Shu, P. S. Yu, and B. Liu, "Understanding Pre-trained BERT for Aspect-based Sentiment Analysis," arXiv preprint arXiv:2011.00169, 2020.

[15] M.P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model," International Journal of Intelligent Networks, vol. 2, pp. 64-69, 2021.

[16] J. Doe, "Generative approach to Aspect Based Sentiment Analysis," Procedia Computer Science, vol. 199, pp. 1-10, 2023. https://www.sciencedirect.com/science/article/pii/S1877050923020203.

[17] A. Silva, Universidade Federal Rural de Pernambuco, Recife, Brazil. https://repository.ufrpe.br/handle/123456789/2133.

[18] P. Anand, "Handle Long Text Corpus for BERT Model," Medium, 15-Jun-2023. https://medium.com/@priyatoshanand/handle-long-text-corpus-for-bert-model-3c85248214aa.

[19] S. Scaria, "Large language models for aspect-based sentiment analysis," arXiv preprint arXiv:2310.18025, 2023. https://arxiv.org/abs/2310.18025

[20] B. Jiang, H. Liu, Q. Ma, H. Yang, and M. Yuan, "All in One: An Empirical Study of GPT for Few-Shot Aspect-Based Sentiment Analysis," in Advances in Computational Intelligence, J. V. De Lira, Ed. Cham: Springer International Publishing, 2024, pp. 315-326.https://link.springer.com/chapter/10.1007/978-3-031-51940-6.

[21] We acknowledge that we have used Grammarly and Quillbot/ChatGPT tools for grammar corrections and rephrasing of sentences.