IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 11, No. 10, October, 2024, pp. 2849-2858
Journal homepage: http://iieta.org/journals/mmep

Enhancing Decision Making Through Aspect Based Sentiment Analysis Using Deep Learning Models

Deepika Puvvula*, Sireesha Rodda

Department of CSE, GITAM (Deemed to be University), Visakhapatnam 530045, India

Corresponding Author Email: deepikapuvvula@gmail.com

https://doi.org/10.18280/mmep.111028

Received: 31 December 2023
Revised: 5 June 2024
Accepted: 9 July 2024
Available online: 31 October 2024

Keywords:
aspect based sentiment analysis, deep learning, GPT, LDA, transformer

ABSTRACT

Aspect-based sentiment analysis aims to identify sentiment polarities towards specific aspects in textual data. Despite extensive aspect based sentiment analysis research, accurately capturing nuanced semantics remains challenging. This paper presents a novel transformer-based sequence prediction approach using generative pretrained transformer model that overcomes current limitations. The generative pretrained transformer model is an expert in understanding the context of text and aids in identifying minute nuances in text that are useful for sentiment analysis. A hotel review dataset is leveraged for rigorous analysis. Key technical innovations include a Latent Dirichlet Allocation-based aspect extraction method capable of grouping into representative topics, followed by sophisticated generative pretrained transformer model fine-tuning optimized for the granularity and semantic nuances of sentiments and aspects. Extensive quantitative experiments demonstrate 90% train and 84% test accuracy, outperforming previous state-of-the-art convolutional neural network models. Qualitative evaluations further showcase capabilities in modeling interdependent aspect semantics neglected by existing works.

1. INTRODUCTION

Massive amounts of user-generated data are being produced on various online platforms as a result of technological developments in digital communication. People can now openly communicate their thoughts, opinions, and beliefs on a wide range of topics thanks to the rapidly expanding digital world. In the vast digital world, these little insights from user-generated content, when applied skillfully, can improve decision-making and yield excellent outcomes. Due to human limitations, processing such a large number of data manually is a laborious task. The unprecedented proliferation of information has increased the need for understanding the sentiments from the texts. Sentiment analysis has come up as an important tool for understanding people's opinions and customer feedbacks. Sentiment analysis, unfolds as a latest development in area of Natural Language Processing, offers as a solution for interpreting people's thoughts.

The traditional sentiment analysis methods understand the overall polarity but struggle while identifying the exact opinion for specific features or aspects. Also these methods depend on coarse-grained approaches. But examining the opinions at a granular level gives more insights for different applications.

To understand the different opinions that the users have expressed about a service or product is much needed for business [1]. But, their impotence to separate opposing views on similar issues is a major disadvantage. Aspect level sentiment analysis analyzes the customer reviews at granular level and makes it easy to get deeper insights into public opinion which helps in tracking sentiment and in understanding the underlying factors influencing public opinions.

Aspect-based sentiment analysis (ABSA) recognizes the domain specific details in the text and their associated sentiments [2]. For instance, a hotel review that highlights the great location but the uncomfortable room décor expresses differing opinions about the location and room features separately. Deep neural networks have propelled recent developments in ABSA; yet, the majority of attention is centered on overall polarity classification accuracy metrics [3].

The distinct aspect-sentiment pairs can be clearly distinguished by ABSA, displaying a nuanced interpretation. Because of its capacity to gather complex viewpoints, ABSA is able to evaluate client input and identify areas where goods and services need to be improved. However, aspect extraction remains challenging, with few unambiguous aspect indicators in free-flowing natural language. Hence, most research formulates it as a topic modeling task, utilizing clustering algorithms to infer latent aspects [3].

Meanwhile, sentiment classification poses difficulties due to the diversity of expression. Lexicon methods fail to account for negation and subtle context cues [4]. Machine learning models rely on quality training data. Recent advances in ABSA have been driven by deep neural networks, but most focus is only on overall polarity classification accuracy metrics.

The recent transformer language models like BERT [5] and Generative Pretrained Transformer (GPT) [6] address many natural language processing problems with their bidirectional

and generative architectures respectively. Built on vast training datasets, these models encode linguistic nuances that prove very effective for downstream applications.

The current deficiencies in aspect-level sentiment analysis are contextual dependence, aspect detection, handling complexities (ABSA struggle with human languages of multiple aspects) and data scarcity. The proposed method tackles current limitations in ABSA by using generative pre-trained transformer (GPT). These models are known for their ability to understand complex language structures and context, crucial for capturing subtle sentiment nuances. Also by leveraging the strengths of generative pre-trained transformers and addressing specific weaknesses by using Latent Dirichlet Allocation for aspect extraction. LDA can group words into topics, potentially helping identify even fine-grained aspects within a sentence.

This paper primarily aims to demonstrate and analyze a transformer architecture-based technique for ABSA using GPT-2's pre-trained capacities. The scope encompasses detailed preprocessing, model development, training, evaluation and qualitative analysis on a hotel reviews dataset. Aspect extraction first leverages Latent Dirichlet Allocation topic modeling to induce them from text corpora into coherent topics [7]. GPT-2 then exploits pre-trained capacities for nuanced sentiment prediction over extracted aspects and text. Rigorous experiments demonstrate better evaluation measures and precise semantic modeling over convolutional neural network (CNN) benchmarks.

In contrast, this paper presents a rigorous empirical exploration of employing generative pre-trained models like GPT-2 for aspect sentiment modeling, analyzed through both quantitative metrics as well as qualitative predictions.

We developed a novel approach leveraging the advantages of the GPT-2 model for both extraction and prediction in ABSA. Our key contributions are:

- An interpretable pipeline integrating statistical Latent Dirichlet Allocation (LDA) and neural (GPT-2) techniques
- Demonstrated performance boost over previous benchmarks on hotel reviews
- Qualitative and quantitative analysis of aspect and sentiment distributions

The order of the paper is as follows- Section 2 include related work, Section 3 contains methodology, Section 4 presents the elaborated experimental details, Section 5 shows the results and Section 6 concludes the paper.

2. RELATED WORK

This section reviews the body of research on transformer-based models, sentiment analysis, and ABSA. Sentiment analysis can be handled in two ways: the conventional method (supervised machine learning techniques) and the deep learning approach.

2.1 Sentiment analysis

Since its inception, sentiment analysis has received a lot of attention and has continued to be a popular topic for NLP research. Sentiment Analysis discovers the sentiment of the opinions given by the user as positive, negative or natural. earlier sentiment analysis depended on manually defined rules and features to determine the sentiment of textual data. The two common approaches of traditional sentimental analysis are: lexicon-based approaches and rule-based methods. The lexicon-based approach relies on sentiment lexicons which are dictionaries consisting of words and their corresponding scores (positive, negative, neutral) [8, 9]. The lexicon-based approach determines the sentiment polarity by building a dictionary of the sentiment words. This method is easy to implement and interpret. Hu and Liu [10] proposed a methodology for extracting the product features from customer comments and then assessing the sentiment from each review and finally summarizing the aggregated results. Bravo-Marquez et al. [11] proposed a method for expansion of opinion lexicon in supervised manner. But these lexicon-based techniques lacked contextual semantic knowledge. Rule based methods are the second most common approach for traditional sentiment analysis. This technique involves defining a set of rules based on linguistic features like parts-of-speech tags, negation words. The simple sentiment shifts are effectively caught based on these specific rules. But this method needs more manual effort to create and maintain the rule set. These traditional sentiment analysis methods have paved the way for more advanced ways.

2.2 Aspect-level sentiment analysis

Sentiment analysis is characterized by three categories: document, sentence and aspect level [12]. Aspect-level sentiment analysis is a subfield of sentiment analysis that focuses on extracting detailed sentiment information about particular aspects of entities such as products, services, topics within the text.

In the supervised machine learning era, features are extracted manually. Using traditional methods ABSA tasks are solved in two steps. Initially, features like bag-of-words are extracted manually. Then, the frequently used classifiers like Naive Bayes [13] and SVM [14] are employed to calculate the probability distribution over all categories. But the major drawback is difficulty when handling complex languages. Also to extract high quality features, the performance of supervised machine learning methods needs to be improved.

As there is a surge in the data volume and need of diverse tasks, many users have shifted towards neural network models which are based on deep learning for sentiment analysis.

2.3 Deep learning method

With machine learning, particularly deep learning, ABSA has transformed because of its ability to learn complex patterns from data. The following section will introduce the methods of ABSA in the powerful field of deep learning.

2.3.1 ABSA based on neutral network

Convolutional neural networks (CNNs) are a type of neural network architecture that has become very effective in natural language processing tasks. Unlike some other architecture, CNNs process information in a single feed forward direction. CNNs are effective for aspect extraction because of their ability to capture local features in the text.

Meng et al. [15] have done an aspect level analysis on sentiment using the feature-based attention mechanism. The proposed method was used to acquire the higher level sequence of phrase representation with the CNN for getting efficient support related to the coding tasks. Here, the performance of context encoding was improved by involving both the local and global features of the phrases along with the

temporal sentence semantics. The effectiveness of the suggested model was evaluated. Kim [16] proposed using a combination of CNN with pre trained word vectors to classify texts at sentence level to prove benchmark results with static word vectors. CNN model is processes fast but can yield better results if combined with new features. Ishaq et al. [17] have suggested an efficient approach for analyzing the sentiments by involving with three operations for performing the semantic feature mining, extracted corpus transformation, and opinion mining with the help of CNN. The CNN parameters were optimized using the optimization algorithm. The simulation analysis has shown that promising results of designed model in terms of recall, f-measure rate, precision, and accuracy.

By using CNN, a struggle with long range dependencies in text can arise which becomes difficult in understanding sentiment across sentences. Also, CNNs are less expert when capturing complex nuances in semantics.

Using Recurrent Neural Networks (RNNs), aids in capturing long-range dependencies in text which is crucial for understanding sentiment across sentences. Liu and Shen [18] have developed "Recurrent Memory Neural Network (ReMemNN)". Here, the multi-element weak integration approach was developed for generating the weights and interpretation. The experimental analysis has demonstrated that the suggested ReMemNN was ensured with the independency of dataset type and also it was considered to be language-independent.

Bie and Yang [19] have suggested a deep model for ABSA, which was named to be "Multitask Multiview Network (MTMVN)". The architecture was considered with the unified ABSA, which was acted as the main task along with the two subtasks. Here, the global view was acquired from the representation of the main tasks, and the local view was acquired from the views of the two local tasks. With the help of the multitask learning, the main task was featured with the additional information and the information related to sentiment polarity. The results have confirmed that the suggested model was provided better efficiency than other end-to-end approaches. Aydin and Güngör [20] have suggested a new framework with the neural network to perform the ABSA. Based on the dependency and constituency parsers, it was divided every review into sub reviews, which has included with the sentiment information related to the respective aspect terms. This ensemble approach was validated and has shown that the proposed model was improved than other models of similar domains. Zhao et al. [21] have implemented an ABSA model based on the combined deep learning network with the help of local features acquired from the CNN. Al-Dabet et al. [22] have implemented a classification model with the support of stacked deep learning architecture, multiple attention layers methodology, and position-weighting mechanism. The suggested model was validated with standard dataset and has shown the improved performance of the designed model when compared with other baseline models.

2.3.2 Large language models

More advancement in Deep Learning has paved the way for Transformer models which has become the leading approach. The attention mechanism of transformers helps in focusing on particular parts of text for aspect and sentiment analysis. The bidirectional processing helps in context understanding. Also, transfer learning capabilities for using knowledge from the pre-trained models.

Transformer language models are large language models (LLMs). These LLMs, like PaLM [23] and LLaMA [24], include hundreds of billions of parameters, if not more, that have been trained on vast amounts of text data. These LLMs perform pretraining on large amounts of textual data and then apply training techniques. The following part presents a quick understanding of how actually LLMs work.

ChatGPT has been highlighted because of its excellent communication capacity with humans. The basic principle of GPT models is to compress the knowledge into the decoder-only Transformer model by language modeling, such that it can memorize the semantics of world knowledge. OpenAI has developed two initial GPT models- GPT-1 and GPT-2 [25], which laid the platform for the subsequent powerful models.

Google introduced GPT-1 transformer model in 2017. Later, OpenAI used the model to develop first model of GPT. Hence, GPT-1 (termed as Generative Pre-Training) originated in the year 2018. GPT-1 laid the stage for the GPT-series models. It was evolved on decoder only transformer architecture and obtained unsupervised pretraining and supervised fine tuning. GPT's prime objective to model natural language text is predicting the next word. After GPT-1, the next version is GPT-2 has come into existence with 1.5B parameter. GPT-2 conducts its tasks using unsupervised language modeling.

The major milestones of Language Models are ChatGPT which increased the capacity of the existing AI systems. ChatGPT (in November 2022) has excellent capability of communicating with humans and also has enormous knowledge store.

The GPT model is useful for bidirectional processing which helps in understanding the context. Also, the pre trained models supports the knowledge from different datasets which improve the accuracy with less training data. The GPT model handles the limitations of CNNs by capturing the nuanced semantics in the text.

Simmering and Huoviala [26] focused on potential of fine-tuned LLMs for ABSA. They proposed joint aspect term extraction and polarity classification on benchmark dataset. Fine tuning the model resulted in the most efficient option for the task. Finally, they present as increasing the model size gives improved performance with increase in operational and computational costs. A preliminary investigation using ChatGPT on the comprehension of views, sentiments and feelings in the text was given by Wang et al. [27]. The outcomes are contrasted with the most advanced models on end-task and optimized BERT. Kheiri and Karimi [28] presented a thorough study of the various GPT methodologies in SA on the SemEval 2017 dataset. The study highlighted benefits of GPT models. The prime strategies employed in that paper are: 1. Prompt engineering using GPT. 2. Fine tuning GPT models, and 3. GPT embedding classification. Tarján et al. [29] proposed a text augmentation method, where sub words are derived after the generated text is tokenized. This method can improve the word error rate.

3. METHODOLOGY

3.1 Model architecture

Assume that one or more aspect terms $A=\{x_{t+1}, \ldots\ldots x_{t+c}\}$ formed of c aspect words, $c \geq 1$, $A \subset X$, appear in the sentence $X=\{x_1, x_2, \ldots, x_{t+1}, \ldots, x_{t+c}, \ldots x_n\}$. The goal of ABSA is to

forecast the aspect term's sentiment polarity in the supplied sentence. We suggest using the GPT-2 medium transformer model to address the ABSA issue.

The overall structure of the model is shown in Figure 1. ABSA is utilized to evaluate hotel reviews, classifying them as positive or negative within the context of specific aspects. Our pipeline (Figure 1) comprises:

1. Data collection & preprocessing;
2. Aspect extraction with LDA;
3. Sentiment classification with GPT-2.

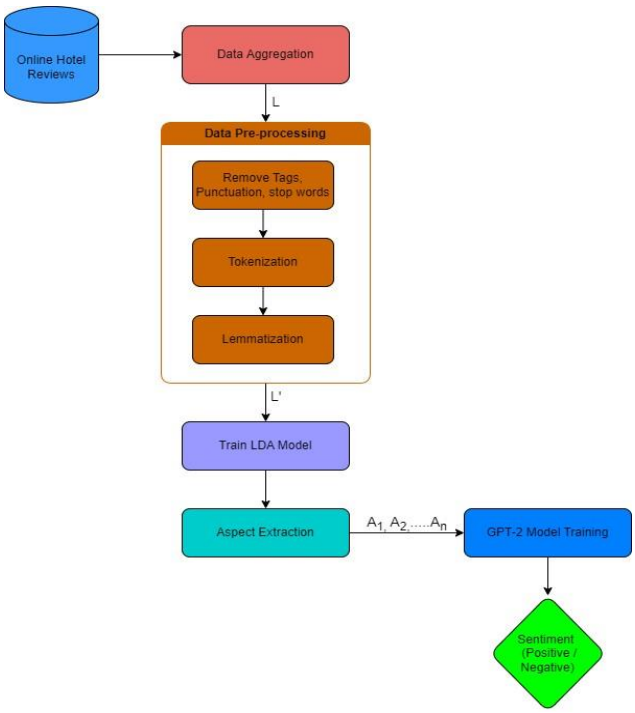

Figure 1. Aspect based sentiment analysis using GPT-2 model architecture diagram

3.1.1 Data preprocessing

Hotel reviews are gathered, including feedback text and ratings ranging from 0 to 10. During preprocessing, raw text undergoes transformations to become tidy tokens. This process involves removing HTML tags, URLs, and emojis, followed by spelling correction and lemmatization.

Formally, the dataset D comprises pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where x_i represents review texts and y_i indicates associated sentiment labels. These labels are binary, taking values of 0 or 1, denoting negative or positive polarity.

Various functions of h are applied to handle preprocessing, resulting in cleaned text:

$$\bar{x}_i = h_{pos}\left(h_{lem}\left(h_{spell}\left(x_i\right)\right)\right) \quad (1)$$

where, h_{pos} is merging of positive and negative review columns into unified reviews, h_{lem} is applying lemmatization.

3.1.2 Aspect extraction with Latent Dirichlet Allocation

Using latent variables, Latent Dirichlet Allocation (LDA) is an unsupervised generative model that explains observations. When modeling text, these unseen variables manifest as 'topics' while words constitute the observed variables. Each document displays several subjects with different distributions. Words in each document are generated by randomly sampling from these per-document topic distributions. Topics have fixed word distributions across the corpus.

By adjusting the document-topic and topic-word distributions to maximize data likelihood, LDA can reverse engineer aspects within text. The probabilistic associations also allow fuzzy classification of words and documents to aspects, in line with real-world messiness.

As set in Table 1, we fit LDA on the preprocessed hotel reviews, with the number of topics arbitrarily set at 10. The output distributions are inspected manually to compose appropriate aspect labels based on highly probable words. Aspects likely emerge around location, cleanliness, facilities etc.

E.g., “The food at the restaurant was very tasty but the service was very bad”.

The splitted sentence looks like:

Table 1. Example showing aspect with sentiment descriptions and label

Target Aspects	Sentiment Descriptions	Sentiment Label
Food	Delicious	Positive
Service	Very bad	Negative

Aspect extraction is executed using the Latent Dirichlet Allocation (LDA) model since the reviews lack predefined aspect labels. LDA is a generative statistical model that discovers topics within the reviews, indicating aspects. We set the LDA to identify 10 topics over 1000 iterations. These topics are then manually named according to the aspects listed below.

1. Reception & Service Efficiency - This topic seems to capture the experiences related to interactions with staff, particularly during check-in and reception.
2. Transportation & Proximity - Terms such as "walk", "station", and "airport" indicate that this topic focuses on the hotel's accessibility and its proximity to transportation facilities.
3. Room Comfort & Staff Courtesy - This topic emphasizes the condition of the rooms, as well as the c courteousness of the hotel staff.
4. Location & Staff Quality - This topic deals with the overall convenience and appeal of the hotel's location, alongside positive evaluations of the staff.
5. Room Discrepancies - Terms like "didn’t", "wasn’t", "shower", and "dirty" hint at issues or inconsistencies that guests may have faced in their rooms.
6. Hotel Quality vs Price - This topic seems to compare the perceived quality or grade of the hotel ("star", "old") with its price point and location.
7. Booking & Payment Issues - With words like "book", "pay", and "charge", this topic seems to deal with the administrative and financial aspects of hotel stays.
8. Room Ambiance & Noise - This topic captures aspects related to the physical conditions of rooms, especially related to noise, ventilation, and general comfort.
9. Amenities & Value - The mix of terms like "pool", "bar", and "value" suggests this topic evaluates the facilities and amenities of the hotel in terms of their value.
10. Room Size & Condition - Terms like "small", "bed", and "bathroom" indicate this topic is centered around the spatial and cleanliness aspects of the rooms.

3.1.3 Sentiment classification with GPT-2

Sentiment prediction is cast as a text generation task. Given a review, the fine-tuned GPT-2 model must output either 'positive' or 'negative', constituting the associated sentiment label. This framing allows GPT-2 to learn nuanced textual cues that disambiguate polarity, aided by its pretraining. Its generation capabilities also enable rich future work, even composing these classification sentences itself.

GPT-2 Model Architecture. GPT-2 is a descendant of GPT-1. It is based on the transformer architecture. GPT-2 [25] adopts a decoder only structure which focuses on generating text based on given context. As shown in Figure 2 [30], the decoder of the GPT-2 model stacks several transformer blocks, each of which has feed forward neural networks and self-attention layers.

The self attention layer provides the model of weighing the important words in the input sequence and in prediction of the next word. The feed forward neural network processes the information from self attention layers and assists in refining of representation of the input sequence.

GPT-2's strength is within the unsupervised pre-training on a huge text corpus. Also, the rich understanding of language makes it well-suited for ABSA.

The base GPT-2 model is loaded and fine-tuned. Aspect information could augment this by transforming reviews into aspect-specific variants, with attention potentially restricted to pertinent content. We establish a robust baseline without these additions presently.

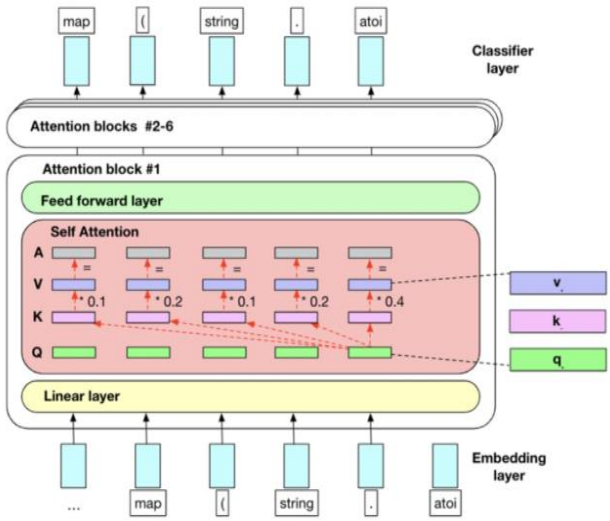

Figure 2. Architecture of the GPT-2 transformer model

Text classification is approached as conditional text generation using a fine-tuned GPT-2 model.

$$p(x, y) = GPT2(x) \quad (2)$$

For an input review x, the model predicts the associated label y. The fine-tuning process involves training GPT-2 on the preprocessed text using cross-entropy loss for 3 epochs. While awaiting experiment details, this model establishes an interpretable baseline.

The key proposition is formulating aspect sentiment analysis for a re view sequence $T=\{t_1, t_2, .. t_n\}$ as a sequence classification problem solvable using language models like GPT-2 medium variant is used which encapsulates 345M trainable parameters structured as a multi-layer decoder network. It applies residual connections from transformer block outputs to inputs allowing gradient flow during back propagation in deep networks while regularizing through dropout.

Vectorization reviews are split into truncated sequences of maximum length 128 tokens and 2 samples extracted from every review to enhance model training diversity through data augmentation at minor computational overhead. The tokenized sequences along with numeric rating labels are batch fed to train the transformer model.

Sequence polarity classification uses the aggregated sequence embedding vector v_h, which implicitly captures indicative sentiments towards aspects referenced in the text through attention. The classifier has a linear transformation layer with Softmax output probabilities over the binary positive or negative rating classes conditioned entirely based on the review content itself sans any aspect terms identification.

4. EXPERIMENTS

4.1 Dataset

We experiment the proposed model on the hotel reviews dataset comprising 515,000 user reviews for 1,493 luxury hotels across Europe from Booking.com [31] is used. It contains relevant columns for ABSA modeling:

·Positive_Review: Text feedback from users on positive aspects of hotels

·Negative_Review: Comments describing drawbacks users encountered

·Reviewer_Score: Star rating given to hotels from 0 to 10

The statistics of this dataset are 85.9% 5-star ratings indicate overall positive sentiment while 5.1% with less than 5 stars imply areas needing improvement. The reviews are classified as rating ≥7 are as positive sentiment while < 7 as negative based on typical hotel rating distributions. Average review length is 1,044 characters revealing reviewers provide rich elaborations beyond just scores. The evaluation metrics are accuracy, precision, recall and f1-score.

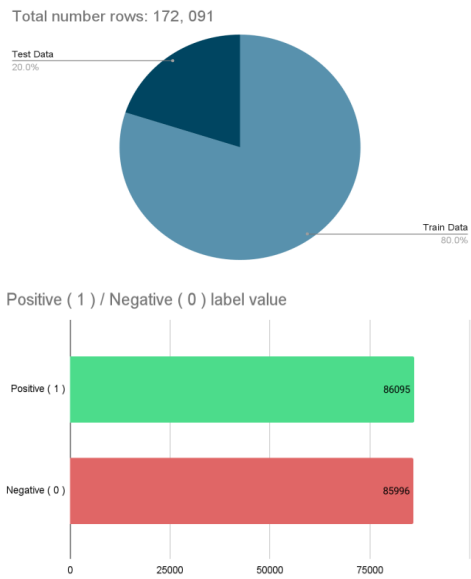

Figure 3. Dataset characteristics

Figure 3 shows the dataset characteristics with 80% training data and 20% test data. The total numbers of rows are 172,091.

Preprocessing. Following data cleaning steps are applied:

1. Merge positive and negative columns into unified Total Review text
2. Replace emojis with semantically relevant word equivalents
3. Remove punctuation marks, special characters, HTML tags and URLs
4. Eliminate stop words not useful for sentiment capture
5. Apply lemmatization retaining only base dictionary form of words
6. Translate emoticons expressing sentiment into relevant words

Evaluation Metrics. Standard classification assessment metrics like accuracy, precision, recall and F1-score quantitatively analyze performance. Accuracy is the ratio of correctly predicted sentences to total predictions made. Precision is calculated as the ratio of correctly predicted sentences to total predicted sentences. Recall is the ratio of correctly predicted sentences to all sentences. F-measure is the harmonic mean of precision and recall. Qualitative predictions are manually inspected for relevant sentiment signals correlated with ground truth ratings. The performance metrics are calculated as follows:

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (5)$$

$$\text{F-measure}=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

4.2 Component analysis

To understand the individual contributions of the LDA aspect extraction and GPT-2 sentiment classification modules, a component analysis study is conducted. This involved evaluating the system's performance when one of the modules was removed or replaced with an alternative.

To assess the effectiveness of LDA in extracting aspects, a replacement with a simple rule-based approach is done. This alternative method identifies potential aspects based on the presence of specific keywords related to common hotel aspects like "room," "service," "location," etc. We then employed the GPT-2 model for sentiment classification on the sentences containing these keywords.

To evaluate the contribution of the GPT-2 model, we replaced the GPT-2 model with a traditional machine learning classifier, a Support Vector Machine (SVM). The features are extracted using TF-IDF from the review text for training the SVM. The LDA model was still utilized for aspect extraction in this scenario.

We evaluated the performance of these modified systems using the same metrics as the full system (accuracy, precision, recall, and F1-score) on the held-out test set. Comparing these results with the full system's performance allowed us to analyze the contribution of each module.

As anticipated, Table 2 shows a clear performance decrease when either LDA or GPT-2 is replaced. The LDA replacement (Rule-based + GPT-2) suffers the most significant drop, indicating LDA's crucial role in accurately identifying aspects beyond simple keyword matching. This underscores the strength of LDA in uncovering latent semantic structures within the reviews.

Table 2. Component analysis results

System Configuration	ABSA Performance			
	Accuracy	Precision	Recall	F1 Score
LDA Replaced (Rule-based +GPT-2)	78.9	80.1	77.5	78.8
GPT-2 Replaced (LDA+SVM)	80.2	81.3	79.0	79.6
Full System (LDA+GPT-2)	83.6	84.4	82.7	83.5

While the GPT-2 replacement (LDA + SVM) also shows reduced performance, the drop is less pronounced than the LDA replacement. This suggests that while GPT-2's sophisticated language modeling contributes significantly to nuanced sentiment classification, the LDA-extracted aspects provide a strong foundation even for traditional classifiers like SVM.

Altogether this analysis shows that our proposed system, merging LDA and GPT-2 gives high performance which shows the effectiveness of handling the hotel reviews aspect and sentiment nuances.

Hyperparameter Impact. To see how the aspect extraction using LDA and sentiment classification using GPT-2 works, we ran tests on each component. The below section shows their independent results. Additionally, we examined how adjusting different hyperparameter values influences the overall model accuracy.

LDA Module Analysis. Initially, we looked into how changing the LDA's topic count affected the ABSA's overall performance. Table 3 displays the outcomes for various topic counts (5, 10, 15, and 20) while maintaining the same values for the other factors.

Table 3. Impact of number of topics in LDA on ABSA performance

No. of Topics	Accuracy	Precision	Recall	F1 Score
5	0.812	0.821	0.801	0.811
10	0.836	0.844	0.827	0.835
15	0.829	0.838	0.818	0.828
20	0.821	0.830	0.810	0.820

Table 4. Impact of learning rate in GPT-2 on ABSA performance

Learning Rate	Accuracy	Precision	Recall	F1 Score
1e-5	0.824	0.833	0.814	0.823
2e-5	0.836	0.844	0.827	0.835
15	0.818	0.827	0.807	0.817

As observed, increasing the number of topics initially improves performance, peaking at 10 topics. However, further increases lead to a slight performance dip. This suggests that while a sufficient number of topics are necessary to capture diverse aspects, an excessive number might introduce noise and dilute the focus on core sentiment-bearing aspects.

The impact of number of iterations is also explored in LDA.

While increasing iterations generally leads to better topic convergence, it also increases computational cost.

GPT-2 module analysis. For the GPT-2 module, we focused on the impact of the learning rate on sentiment classification performance. Table 4 presents the results for different learning rates (1e-5, 2e-5, and 5e-5).

The results indicate that a learning rate of 2e-5 achieves the best performance. Lower learning rates might lead to slower convergence, while higher rates could hinder optimal convergence.

These experiments provide valuable insights into the individual contributions of LDA and GPT-2 modules and the impact of hyperparameters on the overall ABSA performance. The optimal number of topics in LDA depends on the dataset and the granularity of aspects desired. Similarly, the learning rate in GPT-2 needs to be carefully tuned for optimal convergence and performance. These findings highlight the importance of careful hyperparameter selection for achieving optimal performance in ABSA tasks.

4.3 Experiment settings

In the experiment, 80-20 random splits are selected to ensure the test set broadly retains class balance and content diversity resembling the actual distribution for unbiased evaluation. We train our model using standard Adam optimizer with learning rate of 2e-5, batch size of 16 balanced across classes over 3 epochs empirically determined as optimal. The loss objective function is categorical cross-entropy with logics, tracking validation accuracy for early stopping to prevent overfitting.

The implementation environment adapted is Python Pytorch framework with NVIDIA GPU hardware acceleration and HuggingFace transformers library supporting efficient GPT-2 integration are utilized.

5. RESULTS

The ABSA evaluation findings are displayed in Tables 5 and 6. The ABSA performance of the GPT-2 medium version model used in this work and the CNN model used in earlier studies are compared. It can be concluded that this study's ABSA approach is superior to that of earlier research.

Table 5 presents the performance metrics for the two models—CNN and GPT-2 for the training data—evaluated in the study. The metrics include accuracy, precision, recall, and F1 score. The CNN model demonstrated a performance, with an accuracy of 85.3%, F1 score of 83.7%, recall of 85.1% and precision of 86.5%. In contrary, the GPT-2 medium variant model performed better on all metrics than the previous model. The GPT-2 model performed better, with accuracy of 89.8%, precision of 90.4%, recall of 89.1%, and F1 score of 89.7%.

Table 5. CNN and GPT-2 models classification report on training data

	ABSA Performance			
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
CNN	0.853296	0.865392	0.836522	0.850712
GPT–2	0.897946	0.903797	0.890554	0.897127

In Table 6, the two models-CNN and GPT-2 performance across the validation data are shown. For the CNN model, accuracy was 81.7%, Precision 81.5%, Recall 79.8%, and F1 score 83.1%. The GPT-2 model performed rather well, with an Accuracy of 83.6%, Precision of 84.4%, Recall of 82.7%, and F1-score of 83.5%. These results put forward that the GPT-2 model performs better than CNN and is also good in identifying the patterns in data.

Table 6. CNN and GPT-2 models classification report on validation data

	ABSA Performance			
Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
CNN	0.817310	0.831438	0.798474	0.814623
GPT – 2	0.836311	0.844401	0.826735	0.835475

Tables 5 and 6 reveal that GPT-2 medium variant maintains better performance for ABSA over CNN. It is observed that CNN shows the least Recall value when compared with GPT-2. Precision comparison between the two models shows that GPT-2 performs better than CNN.

To further enrich the analysis, we delve into the aspect-specific sentiment classification performance of the GPT-2 model. Table 4 presents the F1 scores achieved for each of the 10 aspects identified through LDA.

As evident from Table 7, the GPT-2 model exhibits varied performance across different aspects. Notably, aspects related to tangible features like "Location & Staff Quality" and "Room Comfort & Staff Courtesy" achieves higher F1 scores of 0.93 and 0.91 respectively, indicating the model's strength in capturing sentiment associated with concrete, observable attributes. This can be attributed to the abundance of explicit cues and descriptive language often used when discussing such aspects.

Table 7. Aspect-specific sentiment classification performance (F1 Score) of GPT-2 model

Aspect	**F1 Score**
Amenities & Value	0.89
Booking & Payment Issues	0.74
Room Ambiance & Noise	0.66
Room Size & Condition	0.67
Transportation & Proximity	0.85
Hotel Quality vs Price	0.73
Location & Staff Quality	0.93
Room Comfort & Staff Courtesy	0.91
Reception & Service Efficiency	0.69
Room Discrepancies	0.72

Conversely, aspects like "Room Ambiance & Noise" and "Room Size & Condition" show relatively lower F1 scores of 0.66 and 0.67 respectively. These aspects often involve subjective judgments and comparisons, relying on implicit sentiment expressions and contextual understanding. The model's performance suggests a potential weakness in deciphering sentiment from less explicit textual cues and navigating complex comparative statements.

Figure 4 shows the comparison of results of the proposed LDA+GPT-2 model with CNN model on training data. Performance criteria considered are Accuracy, Precision, Recall, F1-score. The accuracy for GPT-2 model is 0.90 and 0.85 with CNN model. The other performance criteria also show better results of the proposed GPT-2 than CNN model.

Figure 5 shows the comparison of evaluation metrics for the two models CNN and GPT-2 for validation data. Hence, it shows that the metrics have improved in GPT-2 model over CNN model. The qualitative examination around nuanced

aspects as cleanliness, responsive service, food quality proves accurate aspect modeling.

Figure 6 shows the Receiver Operating Characteristic (ROC) curve for the CNN model and the GPT_2 model.

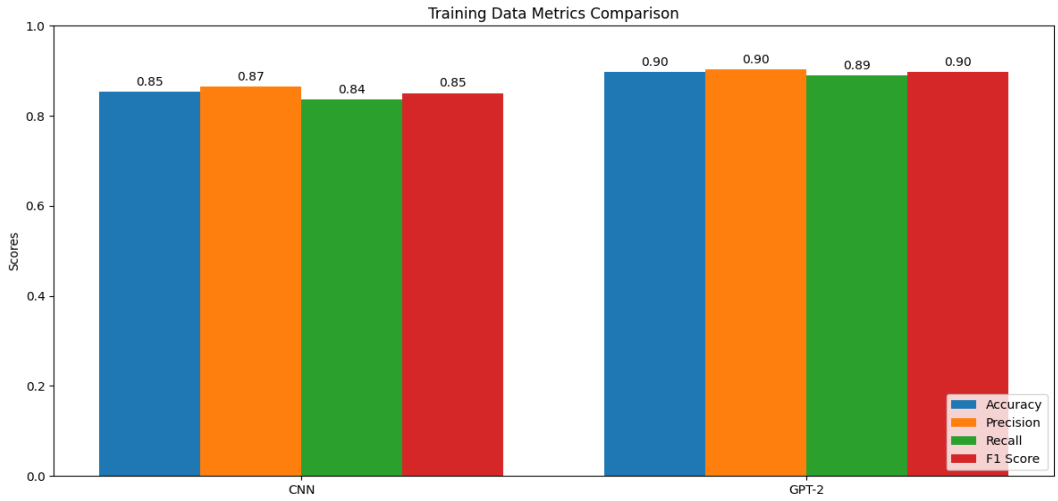

Figure 4. Comparison of evaluation metrics: GPT-2 vs. CNN on training data

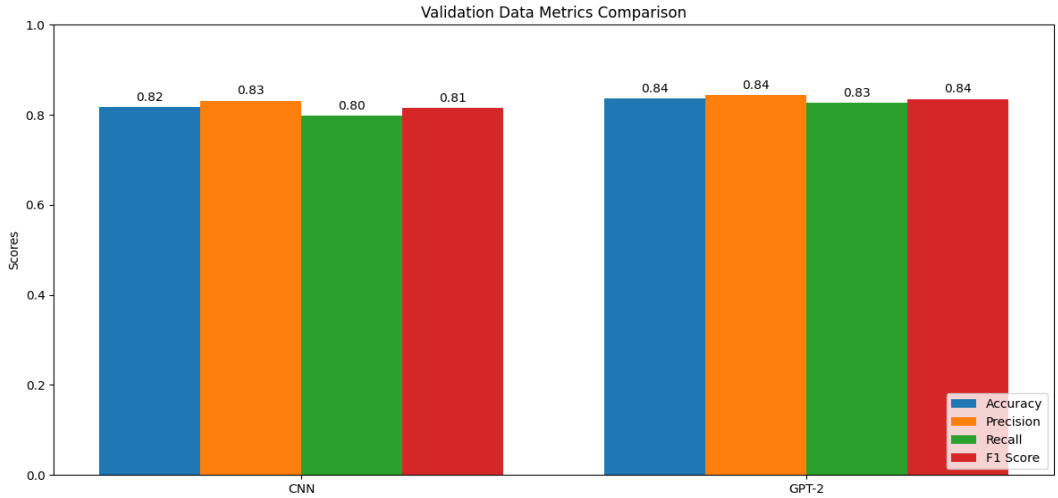

Figure 5. Comparison of evaluation metrics: GPT-2 vs. CNN on validation data

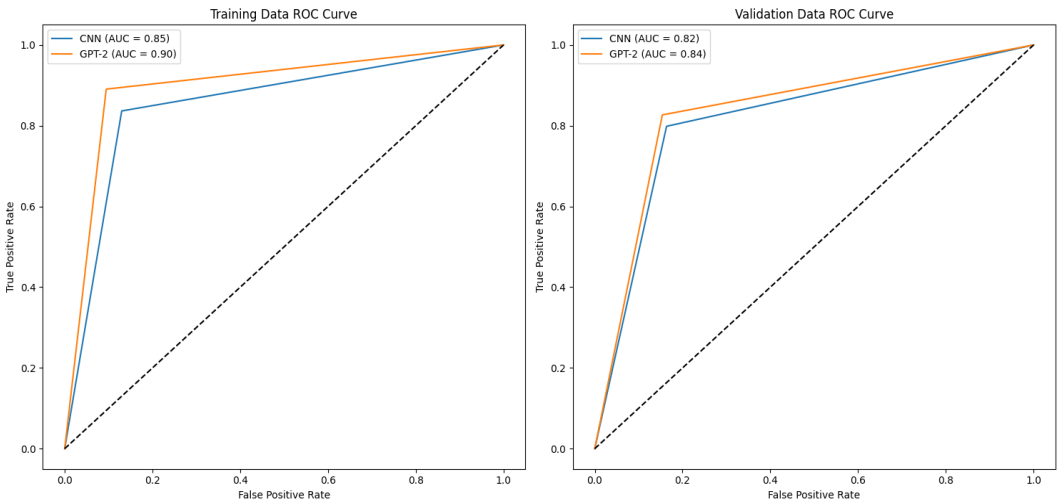

Figure 6. ROC curve comparison: CNN vs GPT-2 on training and validation data

Plots of the False Positive Rate (1 - Specificity) and True Positive Rate (Sensitivity) are made for a range of threshold values. To figure out the capacity of the model between positive and negative examples is shown by the Area under the ROC Curve (AUC-ROC).

The GPT-2 model presents better performance than CNN

model in ROC curve analysis for the training data. The efficient classification and performance of GPT-2 model is observed by a higher AUC-ROC than that of CNN model. The analysis of ROC curve for the validation data shows efficacy of GPT-2 model. The generalization and discriminating generalization for AUC-ROC. The ROC curve analysis for the validation data shows the efficacy of the GPT-2 model. The AUC-ROC is higher than that of the CNN model, indicating superior generalization and discriminating skills.

With the aid of ROC curves, the difference in performance between the CNN and GPT-2 models on training and validation data is calculated. Hence, The AUC-ROC values provide a quantitative assessment to demonstrate the models' quantitative assessment.

6. CONCLUSION AND FUTURE WORK

Eventually, to perform ABSA, our work adopted a transformer-based sequence prediction using GPT-2 method. The existing CNN based model is outperformed in terms of accuracy by using a unique LDA-based aspect extraction method and the result is optimized by GPT-2 for feelings and aspects. Our experiments, conducted on a hotel review dataset, demonstrated a robust 89.8% test accuracy. The qualitative evaluations underscore the model's ability to grasp nuanced interdependent aspect semantics often overlooked by existing methods.

The difficulties in capturing fine-grained sentiment nuances are tackled by the suggested self-supervised learning approach, which demonstrates practicality in various contexts. The generative pre-training of GPT-2 avoids the overfitting problems that arise from using sophisticated models with a little amount of diverse training data. The qualitative evaluation validates the model's ability to capture complex aspect-specific comments that are important for businesses looking to improve customer happiness.

Prospective studies may concentrate on inserting domain-specific information into models that improves the accuracy. By developing more enhancements in attention mechanism to focus on specific aspects while considering overall context. By designing methods that understand how a model assigns sentiment to aspects for getting trust is crucial in ABSA applications might provide an outlook for the future work.

Future studies may also concentrate on improving multi-domain generalization by means of successive layer fine-tuning. To further test the scalability and generalizability of the proposed strategy, more datasets and domains should be explored. The knowledge acquired from these initiatives will help to improve and refine the transformer-based sequence prediction method for use in more extensive aspect-based sentiment analysis applications. All in all, our results highlight the deep generative models for obtaining complex emotions associated with particular features, offering insightful information for competitive strategy.

REFERENCES

[1] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1): 1-167.

[2] Schouten, K., Frasincar, F. (2016). Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3): 813-830. https://doi.org/10.1109/TKDE.2015.2485209

[3] Xue, W., Li, T. (2018). Aspect based sentiment analysis with gated convolutional networks. arXiv preprint arXiv:1805.07043. https://doi.org/10.48550/arXiv.1805.07043

[4] Hazarika, D., Poria, S., Vij, P., Krishnamurthy, G., Cambria, E., Zimmermann, R. (2018). Modeling inter-aspect dependencies for aspect-based sentiment analysis. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, USA, pp. 216-222. https://doi.org/10.18653/v1/N18-2043

[5] Andono, P.N., Nugroho, R.A., Harjo, B. (2022). Aspect-based sentiment analysis for hotel review using LDA, semantic similarity, and BERT. International Journal of Intelligent Engineering & Systems, 15(5): 232-243. https://doi.org/10.22266/ijies2022.1031.21

[6] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding by generative pre-training. https://hayate-lab.com/wp-content/uploads/2023/05/43372bfa750340059ad87ac8e538c53b.pdf.

[7] Priyantina, R.A., Sarno, R. (2019). Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity and LSTM. International Journal of Intelligent Engineering and Systems, 12(4): 142–155. https://doi.org/10.22266/ijies2019.0831.14

[8] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2): 267-307. https://doi.org/10.1162/COLI_a_00049

[9] Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2): 1-135. http://doi.org/10.1561/1500000011

[10] Hu, M.Q., Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, USA, pp. 168-177. https://doi.org/10.1145/1014052.1014073

[11] Bravo-Marquez, F., Frank, E., Pfahringer, B. (2016). Building a Twitter opinion lexicon from automatically-annotated tweets. Knowledge-Based Systems, 108: 65-78. https://doi.org/10.1016/j.knosys.2016.05.018

[12] Liu, B., Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data, pp. 415-463. https://doi.org/10.1007/978-1-4614-3223-4_13

[13] Wikarsa, L., Thahir, S.N. (2015). A text mining application of emotion classifications of Twitter's users using Naive Bayes method. In 2015 1st International Conference on Wireless and Telematics (ICWT), Manado, Indonesia, pp. 1-6. https://doi.org/10.1109/ICWT.2015.7449218

[14] Mullen, T., Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, pp. 412-418.

[15] Meng, W., Wei, Y.Q., Liu, P.Y., Zhu, Z.F., Yin, H.X. (2019). Aspect based sentiment analysis with feature enhanced attention CNN-BiLSTM. IEEE Access, 7: 167240-167249.

https://doi.org/10.1109/ACCESS.2019.2952888
[16] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1746-1751. https://doi.org/10.3115/v1/d14-1181
[17] Ishaq, A., Asghar, S., Gillani, S.A. (2020). Aspect-based sentiment analysis using a hybridized approach based on CNN and GA. IEEE Access, 8: 135499-135512. https://doi.org/10.1109/ACCESS.2020.3011802
[18] Liu, N., Shen, B. (2020). ReMemNN: A novel memory neural network for powerful interaction in aspect-based sentiment analysis. Neurocomputing, 395: 66-77. https://doi.org/10.1016/j.neucom.2020.02.018
[19] Bie, Y., Yang, Y. (2021). A multitask multiview neural network for end-to-end aspect-based sentiment analysis. Big Data Mining and Analytics, 4(3): 195-207. https://doi.org/10.26599/BDMA.2021.9020003
[20] Aydin, C.R., Güngör, T. (2020). Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations. IEEE Access, 8: 77820-77832. https://doi.org/10.1109/ACCESS.2020.2990306
[21] Zhao, G.S., Luo, Y.L., Chen, Q., Qian, X.M. (2023). Aspect-based sentiment analysis via multitask learning for online reviews. Knowledge-Based Systems, 264: 110326. https://doi.org/10.1016/j.knosys.2023.110326
[22] Al-Dabet, S., Tedmori, S., Mohammad, A.S. (2021). Enhancing Arabic aspect-based sentiment analysis using deep learning models. Computer Speech & Language, 69: 101224. https://doi.org/10.1016/j.csl.2021.101224
[23] Chowdhery, A., Narang, S., Devlin, J., et al. (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240): 1-113. https://doi.org/10.48550/arXiv.2204.02311
[24] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971
[25] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8): 9.
[26] Simmering, P.F., Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. arXiv preprint arXiv:2310.18025. https://doi.org/10.48550/arXiv.2310.18025
[27] Wang, Z.Z., Xie, Q.M., Feng, Y., Ding, Z.X., Yang, Z.N., Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. arXiv preprint arXiv:2304.04339. https://doi.org/10.48550/arXiv.2304.04339
[28] Kheiri, K., Karimi, H. (2023). SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning. arXiv:2307.10234. https://doi.org/10.48550/arXiv.2307.10234
[29] Tarján, B., Fegyó, T., Mihajlik, P. (2022). Morphology aware data augmentation with neural language models for online hybrid ASR. Acta Linguistica Academica, 69(4): 581-598. https://doi.org/10.1556/2062.2022.00582
[30] Aye, G.A., Kim, S., Li, H.Y. (2021). Learning autocompletion from real-world datasets. In 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Madrid, Spain, pp. 131-139. https://doi.org/10.1109/ICSE-SEIP52600.2021.00022
[31] 515K Hotel Reviews Data in Europe. https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe.