IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 February 2024, accepted 27 February 2024, date of publication 17 April 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3386969

RESEARCH ARTICLE

Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis

NIMRA MUGHAL[1], GHULAM MUJTABA[1], SARANG SHAIKH[2], AVEENASH KUMAR[3], AND SHER MUHAMMAD DAUDPOTA[4]

[1]Department of Computer Science, Center of Excellence for Robotics, Artificial Intelligence, and Block Chain, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan
[2]Department of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Gjøvik, 7034 Trondheim, Norway
[3]Learners.ai, Toronto, ON M5C 2B5, Canada
[4]Department of Computer Science, Sukkur IBA University, Sukkur, Sindh 65200, Pakistan

Corresponding authors: Nimra Mughal (nimra.cs16@gmail.com) and Sarang Shaikh (sarang.shaikh@ntnu.no)

This work was supported by the National Research Program for Universities (NRPU), Higher Education Commission, Pakistan, under Project 20-14457/NRPU/R&D/HEC/2021-2020.

ABSTRACT Sentiment analysis is essential for comprehending public opinion, particularly when considering e-commerce and the expansion of online businesses. Early approaches treated sentiment analysis as a document or sentence-level classification problem, lacking the ability to capture nuanced opinions about specific aspects. This limitation was addressed by the development of aspect-based sentiment analysis (ABSA), which links sentiment to specific aspects that are mentioned explicitly or implicitly in the review. ABSA is relatively a recent field of sentiment analysis and the existing models for ABSA face three main challenges, including domain-specificity, reliance on labeled data, and a lack of exploration into the potential of newer large language models (LLMs) such as GPT, PaLM, and T5. Leveraging a diverse set of datasets, including DOTSA, MAMS, and SemEval16, we evaluate the performance of prominent models such as ATAE-LSTM, flan-t5-large-absa, DeBERTa, PaLM, and GPT-3.5-Turbo. Our findings reveal nuanced strengths and weaknesses of these models across different domains, with DeBERTa emerging as consistently high-performing and PaLM demonstrating remarkable competitiveness for aspect term sentiment analysis (ATSA) tasks. In addition, the PaLM demonstrates competitive performance for all the domains that were used in the experiments including the restaurant, hotel, books, clothing, and laptop reviews. Notably, the analysis underscores the models' domain sensitivity, shedding light on their varying efficacy for both ATSA and ACSA tasks. These insights contribute to a deeper understanding of model applicability and highlight potential areas for improvement in ABSA research and development.

INDEX TERMS Aspect-based sentiment analysis (ABSA), large language model (LLM), GPT, PaLM, BERT.

I. INTRODUCTION

The ever-expanding influence of e-commerce, driven by the rapid growth of online retail giants like Amazon, Walmart, and Alibaba, has cultivated a thriving ecosystem of customer review platforms encompassing a wide array of services and products [1], [2]. With consumers freely sharing their thoughts and feedback, the digital landscape has witnessed an exponential surge in data generation. This surge primarily comprises unstructured textual customer reviews, posing an intricate challenge for businesses when it comes to analysis [3].

The conventional manual examination of such an extensive dataset is both time-intensive and, given its scale, nearly

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

impossible. Automatic sentiment analysis proves to be a very useful tool in addressing this problem [4]. Sentiment analysis refers to the process of text classification based on its overall emotional tone, negative, positive, or occasionally neutral. It is also known as opinion mining. Based on the subjective expressions and subtle emotional undertones present in the text, these categories have been established [5]. Interestingly, many existing sentiment analysis techniques assume a particular sentence or text to have a constant and consistent sentiment, which is not always the case in real-world scenarios [6]. For instance, consider the following hotel review: ''The hotel staff were extremely welcoming and accommodating, but the room was in dire need of renovation.'' In this case, the sentiment within a single review contains both positive (toward the staff) and negative (toward the room condition), illustrating the complexity of sentiment in real-world texts. Hence, fine-grained-level sentiment analysis is required to cater to such challenges.

Sentiment analysis can be broadly classified into three levels: sentence-level, document-level, and aspect-level analysis [7]. At the document level, the primary focus is on summarizing the sentiments expressed in a review to determine whether they predominantly carry a positive or negative sentiment. In contrast, sentence-level analysis is concerned with evaluating the sentiments expressed in individual sentences. Aspect-based sentiment analysis (ABSA), on the other hand, delves into a fine-grained examination of opinions directed toward specific terms or categories, commonly known as targets [8]. Document-level and Sentence-level sentiment analysis tasks have been a long-established area of research in natural language processing(NLP). With the advent of Deep learning models, such as LSTM [9], GPT-3 [10], and Bidirectional Encoder Representations from Transformers (BERT) [11], these two tasks of sentiment analysis have witnessed remarkable advancements. However, the ABSA is relatively a recent task that has become popular very quickly and is evolving quickly [7].

ABSA is further divided into subtasks such as aspect term sentiment analysis (ATSA), aspect term extraction (ATE), aspect category detection (ACD), and aspect category sentiment analysis (ACSA) [12]. ATSA focuses on predicting the sentiment polarity of specific terms mentioned in a review. On the other hand, ACSA aims at predicting the sentiment of broader categories that may not be explicitly mentioned in the review [13]. For example, let's consider the restaurant review: ''The management was extremely welcoming and accommodating, but the food taste was not up to the mark.'' In this case, ATSA would analyze the sentiment towards the terms ''management'' and ''food taste.'' In contrast, ACSA would aim to predict the sentiment regarding broader categories such as ''food quality'' and ''staff'' which are not explicitly mentioned in the review but can be inferred based on the context. ACSA and ATSA are pivotal tasks in the era of Web 3.0, where online reviews play a vital role in shaping the growth of any company. Hence, companies are more often interested in figuring out the customers' sentiments about targeted terms or categories to improve their services [14]. For example, the restaurant owner may be interested in figuring out the customer's sentiments about the two categories namely, food quality and staff behavior.

ABSA is relatively a recent field that has gained widespread popularity and is undergoing rapid transformations due to many real-world applications [7]. A notable progression in the field pertains to the evolution of utilized datasets. In the early stages of ABSA research, authors frequently engaged in scraping and compiling their datasets from the web [14]. However, a pivotal shift occurred when researchers began adopting standardized datasets, such as Twitter [15], laptops and restaurants datasets from SemEval-2014, SemEval-2015 [16], and SemEval-2016 [17] challenges, and Sentihood [18] dataset. With the rise of Deep learning (DL) and remarkable performance in various natural language processing (NLP), recurrent neural network (RNN) [9], convolution neural networks (CNN) [19], and Transformers-based [20] based models have been widely utilized by researchers for ABSA subtasks.

According to recent surveys conducted in 2022 and 2023 [7], [8], [12], CNN-based models [21], [22], RNN-based models utilizing LSTM [23] and GRU [24], Attention-based LSTM models [25], [26], [27], and transformer-based models [28], [29], [30], [31] are widely employed for aspect-based sentiment analysis on cutting-edge datasets. However, there are three major limitations highlighted by these surveys:

1) Each dataset employs a unique DL model to enhance performance. Hence, DL models tailored for specific domains exhibit limited generalizability across different domains.
2) These models have a high reliance on labeled datasets, and the scarcity of such datasets adversely impacts the performance of DL models in ABSA sub-tasks.
3) These DL models predominantly focus on explicit emotions in text, neglecting the nuanced realm of implicit emotions.

To overcome the above-mentioned limitations, a few studies have been recently published in 2023 that proposed hybrid BERT models. For instance, Mewada and Dewang [32] proposed a hybrid model for ABSA using synthetic attention in a BERT with the incorporation of extreme gradient boosting(SA-BERT-XGBoost) which outperforms base models in terms of accuracy and time efficiency with the highest accuracy of 93.71%) on the restaurant16 [17] dataset. However, it is noteworthy that the SA-BERT-XGBoost does not assess the dependency relationship of sentences, leaving this aspect as a potential avenue for future exploration by researchers in the field. In addition, SA-BERT-XGBoost was trained and tested on restaurant and laptop datasets from SemEval-2015 [16], and SemEval-2016 [17] which can be further extended for multi-domains.

Although BERT models have been extensively studied for ABSA, there has been limited exploration into the potential

of newer generative pre-trained transformers (GPT) [10] and large language models (LLMs) such as ChatGPT-3.5, Bard, and BingChat [33]. The existing models predominantly focus on explicit emotions, overlooking the nuanced realm of implicit emotions.

Hence, to overcome the limitations of existing models, this study explores the potential of LLMs for ABSA subtasks. In addition, we perform a comparative analysis of various state-of-art models with the recently proposed BERT base models. Finally, we execute various analyses on standardized datasets as well as the recently published multi-domain datasets. This research study investigates to answer three research questions:

- **RQ1:** To what extent can available ABSA models generalize their aspect-based sentiment analysis capabilities to unseen aspects or domains?
- **RQ2:** To what extent can pre-trained LLMs (e.g. GPT-3, and PaLM) generalize their ABSA capabilities to unseen aspects or domains?
- **RQ3:** How does the performance of LLMs (e.g. GPT-3, and PaLM), in aspect-based sentiment analysis compare to domain-specific models trained from scratch, considering overall accuracy, F1-score, and computational complexity?

This research study conducts a comprehensive review of sentiment analysis datasets tailored for ABSA tasks, analyzing baseline accuracies achieved across these datasets. Additionally, it undertakes a rigorous evaluation of the latest deep-learning models applied to both ATSA and ACSA tasks, utilizing established benchmark datasets. The study goes further to draw a comparison between the performance of ABSA models and LLMs in terms of accuracy, f1-score, and computational resources. In essence, the methodology outlined in this study serves to identify key opportunities, discuss prerequisites for future advancements, and explore the potential of LLMs that could lead to substantial enhancements in ABSA.

The key contributions of this study include:

- Explanation of benchmark and recently proposed datasets for ABSA tasks, accompanied by reported baseline accuracy.
- Systematic evaluation of recent and state-of-the-art deep learning models for ABSA tasks utilizing benchmark datasets
- Performance comparison between ABSA models and LLMs on the benchmark datasets for ABSA.
- An insightful evaluation highlighting the existing challenges faced by current deep-learning algorithms in addressing ABSA tasks, coupled with a presentation of potential future directions for research and developments.

The subsequent sections of this study follow a defined structure: Section II explores benchmarks ABSA datasets and models in the literature, Section III focuses on dataset acquisition, model selection, and evaluation metrics, Section IV evaluates the results of the discussed models, Section V discussed the key findings and also presents the suggestions and recommendations, and the paper concludes in Section VI along with the future research directions.

II. LITERATURE REVIEW

In this section, we provide a comprehensive literature review on aspect-based sentiment analysis(ABSA). Specifically, we examine the evolution of ABSA systems based on statistical models [34], machine learning models [35], and deep learning models [36]. Furthermore, we investigate the applications of transformer-based models [37], [38] in the context of ABSA. Additionally, we explore the existing large language models (LLMs) that can be utilized for ABSA subtasks namely ATSA and ACSA. By synthesizing the current literature, we identify the research gap and propose future directions to enhance the accuracy, effectiveness, and efficiency of ABSA subtasks with the provision of LLM.

A. EVOLUTION OF ABSA

Early approaches to sentiment analysis treated the task as a sentence-level or document-level classification problem, where the overall sentiment of the entire text or the sentence was determined [39]. However, this approach could not capture nuanced opinions about specific aspects or entities mentioned in the text. ABSA emerged as a response to this limitation, aiming to provide a more detailed understanding of sentiment by associating it with particular aspects or features. ABSA has gained prominence due to its applications in understanding user opinions about products, services, or topics in a more granular manner. The two broader categories of ABSA models are deep-learning (DL) and machine-learning (ML) models, as outlined by Zhou et al. [40] in 2019.

1) ML TECHNIQUES

Machine Learning (ML) falls under the Artificial Intelligence (AI) domain, focusing on the development of statistical models and algorithms. ML empowers computers to learn and make predictions, or decisions without requiring explicit programming [41]. The application of ML techniques marked a significant advancement in ABSA. Supervised learning models, such as Naïve Bayesian [42], Support Vector Machine (SVM) [43], and Artificial Neural Networks (ANN) [44], were employed for ABSA subtasks classification [45]. Feature engineering played a crucial role in these models, with researchers extracting relevant features from the text, such as n-grams, bag-of-words, part-of-speech, syntactic structures, and sentiment lexicons [46]. The performance of these methods heavily relies on manually crafted features, which unfortunately are labor-intensive and comparatively less effective. Hence, researchers explored revolutionary DL techniques for ABSA subtasks in recent years.

2) DL TECHNIQUES

Propelled by rapid advancements in neural network techniques, Deep Neural Networks (DNNs) have achieved

noticeable success in various applications. This success has led ABSA research to transition from feature-based techniques to DNN methods. The rise of DNN architectures, such as recurrent neural networks (RNNs) [9], convolutional neural networks (CNNs) [19], and transformers [20] brought about a paradigm shift in ABSA. Wang et al. [22] proposed a unified position-aware convolutional neural network (UP-CNN) for ATSA and ACSA tasks. This study utilized the Laptops and Restaurants datasets from SemEval-2014-Task-4 [16], MAMS-Term [13], and Twitter [15] datasets for ATSA. Whereas, for the ACSA task, Restaurant-14, and Restaurant-large were utilized from the SemEval challenges from the year 2014-2016, and the MAMS-Category [13] dataset. In addition to that, an RNN-based model that involves long short-term memory (LSTM) was proposed by [23], and a gated recurrent unit (GRU) was also proposed by [24]. These models have shown enhanced performance in capturing contextual information and learning intricate patterns within textual data [47]. Attention-based LSTM and CNN models have been widely proposed for ABSA. For instance, Sadr et al. [21] proposed the attention-based CNN model. and attention-based LSTM models were proposed by [25], [26], and [27].

The ability of neural networks to automatically learn hierarchical representations contributed to their success in ABSA subtasks. This shift from traditional feature-based methods to diverse DNN architectures reflects the substantial progress in neural network research and its applicability to ABSA subtasks. Each of these DNN-based approaches brings unique strengths to the table, contributing to the evolving landscape of ABSA. However, there were two main limitations of the LSTM and CNN models. First, the model trained with one dataset was not performing well for other datasets and domains. Secondly, these models were struggling to achieve a remarkable performance in terms of accuracy. Hence, transformer-based models have emerged in recent years as proposed solutions for ABSA tasks.

B. TRANSFORMER-BASED MODELS

The development of transformer architectures, as demonstrated by models such as BERT (Bidirectional Encoder Representations from Transformers), was another significant advancement in ABSA. Transformer models performed exceptionally well in sentiment analysis and other natural language processing tasks, outperforming earlier methods in the capture of long-range dependencies and contextual information [20]. Fine-tuning pre-trained transformer models for aspect-based sentiment analysis tasks became a common practice, allowing models to leverage large-scale pre-training on diverse textual data [48].

Transformers possess a self-attention mechanism that allows them to recognize word dependencies within a text. Unlike recurrent neural networks (RNNs), which process sequences sequentially, transformers can process words in parallel. The self-attention mechanism can be represented by

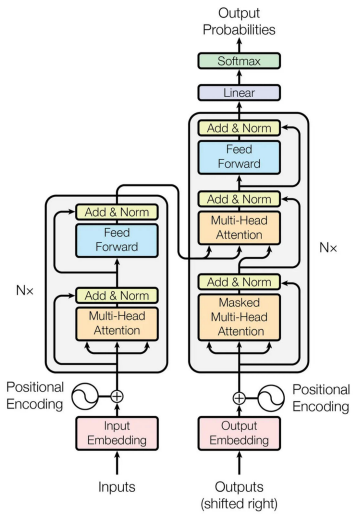

FIGURE 1. Transformer architecture [49].

the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

In this case, the matrix representing the query vectors is labeled as Q, the key vectors are denoted by matrix K, and the value vectors are expressed through matrix V. The similarity between query-key pairs is determined by the product of Q and K^T. This factor is scaled by the $\sqrt{d_k}$ factor, which is the square root of the dimensionality (d_k) of the query/key vectors. The attention output is obtained by multiplying the attention weights element-wise with V after the softmax function has been applied [49].

Figure 1 illustrates how an encoder and a decoder make up a transformer's architecture. A series of symbol representations $(x_1,\cdots,x_n)$ is mapped by the encoder to a series of continuous representations $(z=(z_1,\cdots,z_n))$. Once z is provided, the decoder proceeds to create a symbol output sequence $(y_1,\cdots,y_m)$ by producing each element incrementally [49]. In the context of ABSA, the encoder takes the review and aspect as input and encodes it into a high-dimensional representation. On the other hand, the decoder produces the output, such as sentiment polarity. The self-attention mechanism of transformers enables both the encoder and the decoder to record global dependencies for the review and facilitate effective information flow [20].

1) BERT

BERT (Transformers Bidirectional Encoder Representations) is a powerful transformer-based model that has revolutionized natural language processing tasks. It includes bi-directional context understanding through the Transformer architecture [50]. BERT has achieved exceptional performance across various domains of NLP such as sentiment analysis, text classification, question answering, and machine translation. BERT's contextual understanding of language has led to improved search engines, chatbots, and virtual assistants,

enabling more accurate responses and better user experiences [11]. Although BERT-based models have shown remarkable success in ABSA, challenges persist, including the need for domain-specific pre-training, handling ambiguous contexts, and addressing the scarcity of labeled data for specific domains. Hence, to overcome these limitations, large language models (LLMs) such as PaLM API, GPT-3, and flan-t5 can be used for ABSA subtasks.

2) GPT

GPT is a state-of-the-art LLM and probably a watershed moment in the natural language processing (NLP) field [51]. Developed by OpenAI, GPT employs a transformer architecture to generate coherent and contextually accurate text. GPT is capable of producing fluent text outputs such as language translation, text generation, and question answering in a human-like manner as it has been refined by using a large text data corpus for various NLP tasks [10]. GPT-1, GPT-2 [52], and GPT-3 [51] have been made available by OpenAI.[1] Moreover, the latest innovation of ChatGPT[2] stunned everyone with its sophisticated features and quickly rose to the top of social media and news outlets. The ChatGPT is uniquely capable of accomplishing diverse tasks from the customer reviews datasets such as sentiment analysis, text summarization, text classification, offensive word detection, etc [10], [53]. While ChatGPT has demonstrated remarkable language generation capabilities and has been applied to various NLP tasks, its implementation for ABSA remains limited and requires further scientific inquiry.

There are very few studies in the literature that explore the potential of GPT and other LLMs for ABSA and its subtasks. For example, Chumakov et al. [54] explores the potential of GPT for the Aspect Sentiment Triplet Extraction (ASTE) subtask. Additionally, the article only performed experiments with the SemEval-2016 [17] for Restaurant and Laptop domains. However, the detailed comparative analysis from various domains and the latest LLMs are not presented. Similarly, Another study [55] explores the GPT to predict the sentiment scores of various aspects from the Bakery reviews domain. However, the main purpose of study is to predict the rating score of the bakery by analyzing the sentiment scores of different aspects from bakery domains only. In addition, the comparative analysis with other LLMs is not presented in this study.

To the best of our knowledge, no comparative study has been conducted that explores the potential of GPT and other LLMs for ABSA tasks with state-of-the-art ABSA datasets from different domains. The latest LLM models have not yet been explored thoroughly in the ABSA domain in the above-mentioned studies. Hence, our study is a first attempt that explore the ability of LLMs for the ABSA tasks.

[1]https://openai.com/
[2]https://chat.openai.com/

III. METHODOLOGY

In this section, we discuss the methodology employed to systematically evaluate the performance of LLMs for two main tasks of ABSA. Four main steps are carried out on the acquired datasets, as depicted in Figure 2. First, we acquire the state-of-the-art datasets for two subtasks of ABSA namely, Aspect Term Sentiment Analysis (ATSA) and Aspect Category Sentiment Analysis (ACSA). Secondly, we trained the state-of-the-art LSTM-based models on the acquired dataset to get the baseline results for these datasets. Next, we evaluated the results of BERT-based ABSA models that were recently proposed in the literature. Afterward, we designed a prompt for the GPT-3.5-turbo[3] and Google Bard[4] model through prompt-engineering methods to evaluate the potential of LLMs for ACSA and ATSA tasks. Finally, we conducted a comparison between the baseline results obtained from LSTM-based models and the best results achieved by recently proposed ABSA models and LLMs available to date. The subsequent sections present a detailed description of the methodology from data acquisition to results analysis.

A. TASK DESCRIPTION

This study is focused on two main subtasks of ABSA, namely, aspect category sentiment analysis (ACSA) and aspect term sentiment analysis (ATSA) [13]. ATSA is defined as predicting the sentiment of any aspect term A_t that is part of the review. Given a review $R = \{w_{s1}, \ldots, w_{sn}\}$, an aspect term $A_t = \{w^1_{a1}, \ldots, w^m_{a1}\}$, ATSA aims at predicting the sentiment polarity $y \in \{1, \ldots, C\}$ of the A_t in the review R. In contrast, ACSA is defined as predicting the sentiment of any aspect category A_c that is Entity-Attribute pair from the predefined Entities and Attributes. For instance, the categories for the restaurant dataset can be food-quality, food-price, food-taste etc. Given a review $R = \{w_{s1}, \ldots, w_{sn}\}$, an aspect category A_c, ACSA aims to make predictions regarding the sentiment polarity $y \in \{1, \ldots, C\}$ of review R with respect to A_c

B. DATASET ACQUISITION

In this study, we used state-of-the-art customer review datasets for ABSA subtasks that are publicly available and widely used for ABSA subtasks. In addition, we also used the recently published dataset in the field of ABSA. In total, five different datasets are used to evaluate the performance of LLMs for ABSA subtasks. We followed the training and test splits recommended by the baseline studies as shown in 1. However, in certain datasets, no distinct validation set was provided. In these cases, we adopted a common approach: allocating 20% of the training data for validation, with the remaining portion utilized for actual training. The subsequent sections describe the details of these datasets.

[3]https://platform.openai.com/docs/models
[4]https://bard.google.com/

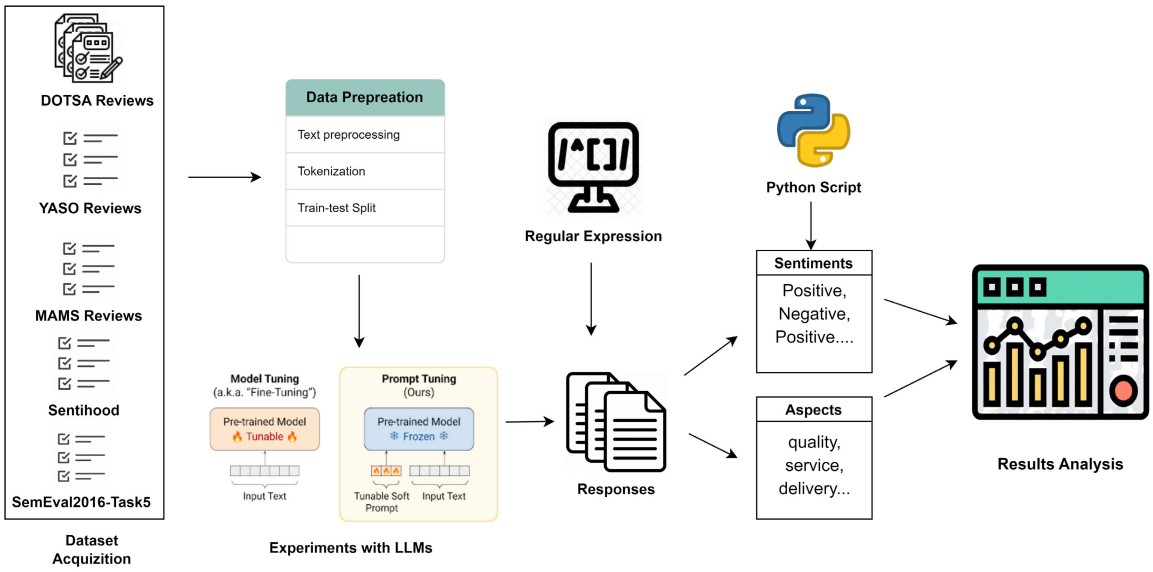

FIGURE 2. Methodology for LLM evaluation for ABSA subtasks.

TABLE 1. Summary of benchmark user review Datasets for ABSA tasks.

Dataset	Year	Domain	Tasks	Train	Val	Test
SemEval2016[17]	2016	Restaurant	ACSA	2507	-	859
		Restaurant	ATSA	1880	-	650
		Laptop	ACSA	2909	-	801
Sentihood [18]	2016	Neighbourhood	ACSA	3401	840	1679
MAMS[13]	2019	Restaurant	ACSA	7090	888	901
		Restaurant	ATSA	11186	1332	1336
YASO [56]	2022	Multi-domain	ATSA	1302	0	332
DOTSA[57]	2022	Books	ATSA	1711	342	368
		Clothing	ATSA	942	177	196
		Restaurant	ATSA	3330	712	682
		Hotel	ATSA	2373	498	549

1) SEMEVAL2016-TASK5

The SemEval-2016 Task5 is a shared task on ABSA which is further divided into subtasks. The datasets were divided into subtasks, including sentence-level and text-level ABSA, and annotated with aspects, opinions, and polarities [17]. However, this study is focused on two subtasks; subtask 2: Aspect Term Polarity (ATP) or ATSA - determining the polarity/sentiment of each aspect term [58], subtask 3: Aspect Category Polarity (ACP) or ACSA - determining the polarity/sentiment of each aspect category in a given sentence or text [59], [60]. The dataset includes training and testing sets for 8 languages and 7 domains, consisting of 19 training datasets and 20 testing datasets. Additionally, a standardized evaluation procedure is provided.

The datasets were provided in various languages, including English, Dutch, French, Russian, Spanish, Turkish, and Chinese. The domains of the datasets included restaurants, hotels, consumer electronics, laptops, and museums. However, this study is focused on identifying the sentiment of the provided term or category from English sentences. Hence, we used the datasets of restaurants and laptop domains that are based on the English language. The dataset for the restaurants' domain consisted of 440 review texts (2675 sentences). In the context of the ACSA task, the dataset comprises a total of 3366 E#A, polarity tuples that have been manually annotated. The category of aspect term is denoted by E#A (Entity#Aspect) pair where E and A must be chosen from predefined inventories of entity types (such as ''restaurant'', or ''food'') and attribute labels (such as ''price'', or ''quality'').

The train test split for these tuples is shown in table 1, 2507 training tuples, and 859 test tuples. Similarly, in the context of the ATSA task, the dataset contains 2530 manually annotated OT, polarity tuples. Here OT represents the opinion terms, which are linguistic expressions found in the provided text referring to the reviewed E#A pair. The train/test split for these tuples are shown in table 1, 1880 train tuples, and 650 test tuples.

2) SENTIHOOD

The SentiHood dataset is designed for targeted aspect-based sentiment analysis specifically focused on urban neighborhoods. It was created and provided by [18]. The dataset was derived from a question-answering (QA) platform where users engage in discussions about urban neighborhoods. It comprises units of text that frequently reference multiple aspects of one or more neighborhoods. The defined aspects in the SentiHood dataset encompass general attributes, price, safety, transit location, tourist spots, nightlife, shopping, and dining. It contains annotations for sentiment polarity and aspect category for each unit of text. It is the first dataset to use a generic social media platform, specifically a QA platform, for detailed opinion mining. The dataset has

been used in various research studies related to aspect-based sentiment analysis and opinion mining [61], [62]

3) MAMS

The MAMS dataset is a challenging dataset for ABSA, in which each sentence contains at least two different aspects with distinct sentiment polarities. It is designed to handle aspect-based sentiment analysis tasks that involve multiple aspects and sentiments in a single sentence. The dataset comprises two versions: one for aspect-category sentiment analysis (ACSA) and another designed for aspect-term sentiment analysis (ATSA). The dataset is available on GitHub,[5] and it has been used in research studies such as "A Challenge Dataset and Effective Models for aspect-based sentiment analysis".[6] The dataset is written in a "standard" (flexible structural) model and has a separation between model code and database. The statistics of the MAMS dataset are presented in 1.

4) YASO

YASO is a targeted sentiment analysis dataset for evaluating aspect-based sentiment polarities in open-domain reviews [56], [63]. It is a crowd-sourced dataset comprising over 2,000 English user comments extracted from Yelp,[7] Amazon [64], Stanford Sentiment Treebank (SST) [65], and OPINOSIS [66]. The dataset comprises 2215 annotated sentences, comparable in size to existing test sets. It encompasses all potential targets, not solely valid ones, each annotated with its confidence level, sentiment label (including raw annotation counts), and span. The YASO dataset is not restricted to any particular review domain, thus providing a broader perspective for open-domain targeted sentiment analysis (TSA). However, some major domains were identified by the authors in [56], as follows: 400 sentences were categorized as restaurants, 412 as electronics, 161 as hotels, 144 as automotive, 500 as movies, and the remaining 596) sentences were labeled as others. Whereas, the domain label is not publicly available with the dataset.

5) DOTSA

DOTSA is a diverse dataset that encompasses customer reviews from six distinct domains, including clothing, book, hotel, restaurant, social media data, and financial news [57]. The dataset sources are as follows:

- Books and Clothing: The dataset includes 928 clothing reviews and 986 book reviews, randomly chosen from the 5-core version of a publicly available dataset.
- Restaurant: Restaurant reviews were collected in Boston from Yelp (as of April 17, 2021). The dataset comprises 940 reviews, specifically selected from restaurant-related content.

[5]https://github.com/siat-nlp/MAMS-for-ABSA
[6]https://aclanthology.org/D19-1654/
[7]https://www.yelp.com/dataset

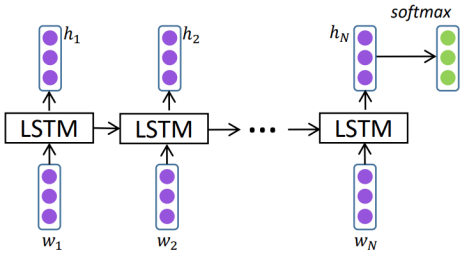

FIGURE 3. Architecture of standard LSTM [68].

- Hotels: Hotel reviews from Boston, collected via Airbnb (as of February 19, 2021), contributed 1029 reviews selected at random.
- Social Media: A subset of 1194 sentences was selected from social media data, and annotators assessed sentiment from an investor's perspective.
- Business News: The business news dataset consists of 936 news articles collected from Reuters and Bloomberg. Reuters News data was collected from March 2021 to April 2021, generating 498 instances, while Bloomberg News data was collected from October 2006 to November 2013, resulting in 438 samples.

C. STATE-OF-THE-ART LSTM-BASED MODELS

The Long Short-Term Memory (LSTM) Network, developed by Hochreiter and Schmidhuber [67] in 1997, excels in comparison to traditional feed-forward neural networks. It adeptly handles both individual data points and sequences due to its incorporation of feedback connections. The LSTM architecture consists of three gates (input, forget, and output) and a cell memory state.

In summary, the LSTM cell computations are expressed as:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Here, i_t is the input gate, g_t is the cell input, o_t is the output gate, f_t is the forget gate, c_t is the cell state, and h_t is the hidden state at time t. The sigmoid function σ is applied element-wise, $\odot$ denotes element-wise multiplication, and the weights and biases are represented accordingly.

In this study, we used three variations of LSTM-based models to establish the baseline results for all the datasets namely, LSTM with aspect embedding (AE-LSTM) [68], Attention-based LSTM (AT-LSTM) [69], Attention-based LSTM with Aspect embedding(ATAE-LSTM) [69]. These LSTM-based models are widely used in the domain of ABSA [12]. The subsequent sections describe the architecture of these models.

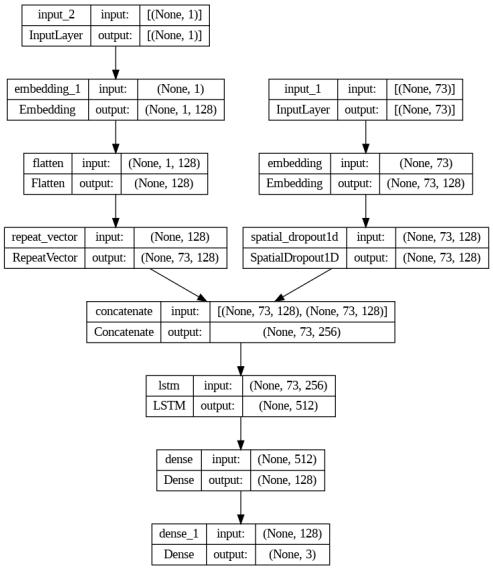

FIGURE 4. Architecture for AE-LSTM.

1) AE-LSTM

AE-LSTM (Aspect-Enhanced Long Short-Term Memory) was proposed by [68]. This is a neural network model designed for aspect-based sentiment analysis, a task involving the classification of sentiment polarity in a text based on a specified aspect or feature. The model leverages two input layers for the main text sequence and the identified aspect, employing word embeddings and a spatial dropout to capture textual information as shown in Figure 4. The aspect is similarly embedded and broadcast across the text sequence for effective fusion. The concatenated embeddings undergo processing by an LSTM layer, enabling the model to capture sequential patterns. A subsequent dense layer with ReLU activation is employed for non-linear feature extraction, culminating in a final dense layer with softmax activation for sentiment classification. This model is particularly promising for nuanced sentiment analysis, allowing for a more granular understanding of sentiment variations across different aspects within a given text.

2) AT-LSTM

The AT-LSTM (Attention-based Long Short-Term Memory) neural network architecture proposed here is specifically designed for aspect-based sentiment analysis, aiming to classify sentiment polarity in a given text concerning a specified aspect or feature [68]. The model incorporates two input layers: one for the main text sequence (input-text) and the other for the identified aspect (input-aspect). Word embeddings are applied to the text sequence using spatial dropout, and a similar embedding process is employed for the aspect. The aspect information is then broadcasted across the text sequence through flattening and repetition operations. Subsequently, an LSTM layer processes the text embeddings, generating hidden vectors for each word in the sequence as shown in Figure 5. The attention mechanism is introduced

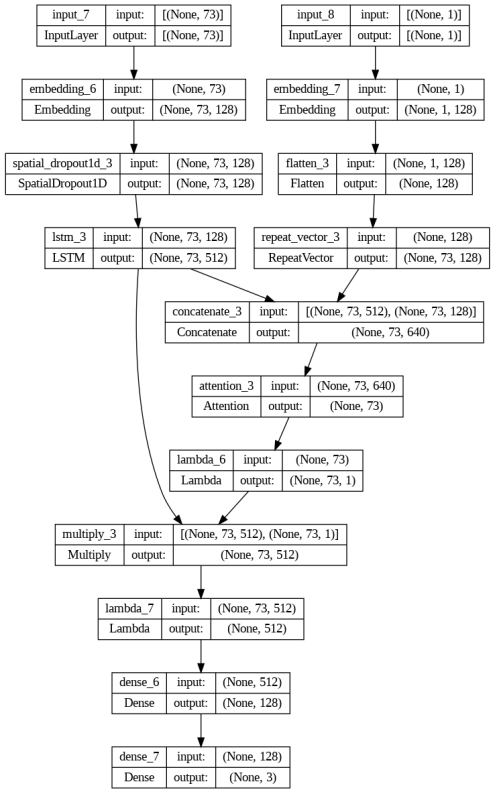

FIGURE 5. Architecture for AT-LSTM.

through the concatenation of hidden vectors and repeated aspect embeddings. The attention weights are computed, enabling the model to focus on relevant parts of the input sequence. The attended hidden vectors are then aggregated, and a dense layer with ReLU activation captures non-linear patterns. The final output layer, with softmax activation, provides the binary sentiment classification. This AT-LSTM model holds promise for aspect-aware sentiment analysis, offering an interpretable approach by emphasizing crucial aspects within the input text.

3) ATAE-LSTM

The ATAE-LSTM (Attention-based LSTM with Aspect Embedding) neural network architecture presented here is tailored for aspect-based sentiment analysis, a task focused on discerning sentiment polarity in a given text concerning a designated aspect or feature [68]. The model comprises two input layers: input-text for the main text sequence and input-aspect for the aspect of interest. Word embeddings are employed to represent words in the text sequence, with spatial dropout applied to prevent overfitting. Aspect embeddings are similarly utilized. The aspect information is then broadcasted across the text sequence through flattening and repetition operations. The concatenated input is processed by an LSTM layer as shown in Figure 6, producing hidden vectors for each word and capturing the final hidden state. An attention mechanism is introduced by concatenating the hidden vectors and repeated aspect embeddings, generating

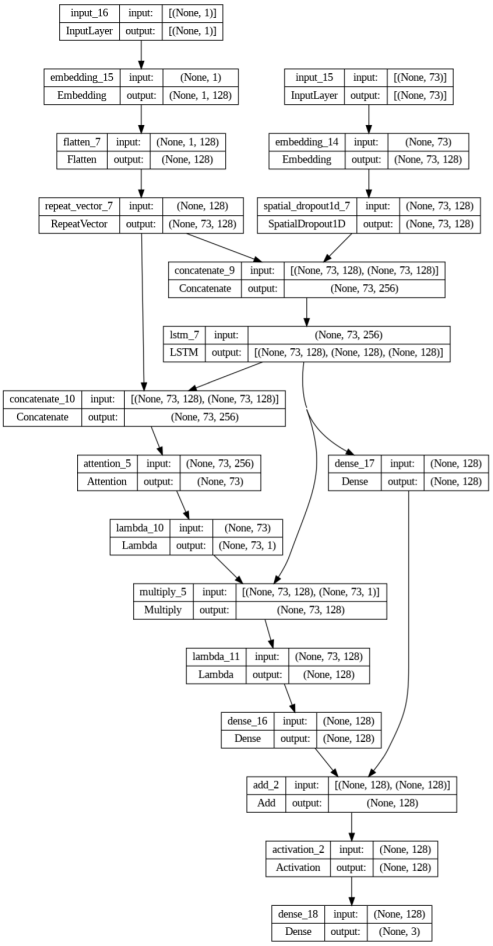

FIGURE 6. Architecture for ATAE-LSTM.

attention weights that emphasize relevant portions of the sequence. The attended hidden vectors are aggregated and processed through dense layers with hyperbolic tangent activation, offering an interpretable approach to sentiment analysis by incorporating specific aspects within the input text. The final output layer, employing softmax activation, provides a binary sentiment classification, reflecting the probability distribution over positive and negative sentiments.

D. FINE-TUNED ABSA MODELS

In this study, have used two widely known models for ABSA subtasks namely, deberta-v3-base-ABSA-v1.1,[8] and flan-t5-large-ABSA.[9] These two models have recently been proposed in the domain of ABSA and are widely used by many researchers for ABSA subtasks. The subsequent sections describe the details of these two models.

1) DeBERTa-v3-BASE-ABSA-v1.1

DeBERTa-v3-base-ABSA-v1.1[10] is a model for aspect-based sentiment analysis (ABSA) that is trained with English

[8]https://huggingface.co/yangheng/deberta-v3-base-ABSA-v1.1
[9]https://huggingface.co/shorthillsai/flan-t5-large-ABSA
[10]https://github.com/microsoft/DeBERTa

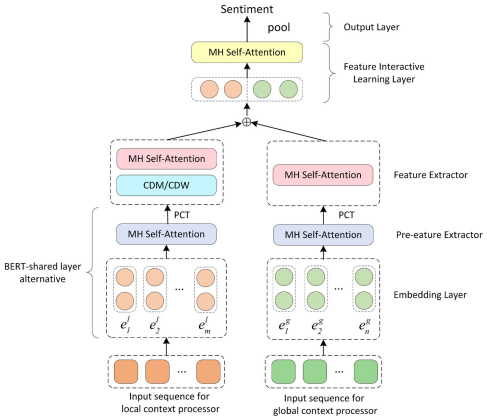

FIGURE 7. Architecture of LCF-BERT. MH self-attention: Multi-head self-attention.

datasets from ABSA Datasets.[11] It relies on the FAST-LCF-BERT model incorporating Microsoft/DeBERTa-v3-base, sourced from PyABSA [70], [71]. The model is trained with 30k+ ABSA samples and fine-tuned with 180k examples for the ABSA dataset including SemEval-2014 Task 4 (Laptop14 and Restaurant14) [72], SemEval-2016 Task 5(Restaurant-16) [17], MAMS [13], Television [73], and TShirt [74], [75]. However, the base model, LCF-BERT is an ABSA model that enhances sentiment polarity predictions by focusing more on local context words through the use of a Local Context Focus (LCF) mechanism [76]. The main architecture of LCF-BERT is shown in Figure 7. There are three primary components of this model:

1) Local Context Focus (LCF) Mechanism: It forms the basis of the LCF-BERT model. The Context Features Dynamic Weighted (CDW) and Context Features Dynamic Mask (CDM) layers are utilized to dynamically modify the context word weights according to their significance to the aspect. This guarantees that the model concentrates more on the words that are most likely to affect the aspect's sentiment polarity.
2) BERT-shared Layer: Long-term dependencies between local and global contexts are captured using the BERT-shared layer. By extracting contextual representations of words using the pre-trained BERT parameters, the model can comprehend the relationships between words throughout the entire sentence or document.
3) Aspect Sentiment Prediction: This part predicts the sentiment polarity of an aspect by using the BERT-shared layer's output and the LCF mechanism's output as inputs. A classification layer is utilized to designate a sentiment label (e.g. positive, neutral, or negative) to the aspect.

2) FLAN-T5-LARGE-ABSA

FLAN-T5-large-ABSA is a fine-tuned version of FLAN-T5-large[12] on a custom dataset prepared by GPT-4 and verified

[11]https://github.com/yangheng95/ABSADatasets
[12]https://huggingface.co/google/flan-t5-large

by a human. flan-t5-large was developed by Google AI and was first introduced in the paper "Scaling Instruction-Finetuned Language Models" [77]. The paper demonstrates that FLAN-T5-large can attain cutting-edge performance on various NLP tasks, such as question answering, natural language inference, and summarization.

The following are some of flan-t5-large's main characteristics:

1) Fine-tuning instructions: FLAN-T5-large was trained on a set of instructions to improve its comprehension and compliance with them.
2) Large size: With 780M parameters, the FLAN-T5-large language model is large and capable of capturing more intricate word-sentence relationships.

E. LLMs

In this study, we have utilized two main LLM models that are widely used and very famous, namely GPT-3.5-turbo[13] and PaLM bison-001 model.[14] GPT-3.5-turbo was released on 23 March 2023 by OpenAI[15] community. This is the latest and most powerful model of GPT that allows access through API keys.

In addition, we used the text-bison-001 model by Google-generativeai and makersuite.[16] By utilizing these models, we predicted the sentiment/polarity for each aspect category or aspect term for the provided review to further enhance the effectiveness of LLMs for ACSA and ATSA tasks. API requests were sent with parameters and prompt engineering using Python 3.9. The cost for using OpenAI's API is $0.002 per 1000 tokens. However, the web version of the same GPT model is completely free. In contrast, google's PaLM API is free as of now.

1) PROMPT DESIGN

While writing the GPT prompt, we followed the prompt engineering techniques by writing a structured prompt for ABSA subtasks. The prompt accepts the review and a list of aspect categories and aspect terms, and it provides the sentiment of each provided sentiment into three classes, i.e. positive, negative, and neutral. The prompt structure is shown in Figure 8.

Prompt Example

As an aspect-based sentiment analyzer, your task is to extract sentiments (Positive, Negative, or Neutral) from people's review for a given list of aspects.

Each review consists of a text and may contain single or multiple aspects and their corresponding sentiments. You will be provided with the review and the list of aspects.

Your goal is to process the reviews and given list of aspects. Then, predict sentiment of each given aspect.

Please present the result in specified format.

Review: {text}
Aspects: {aspects}
Required output Format:
['sentiment1', 'sentiment2', 'sentiment3', ..]

FIGURE 8. Prompt used in this study.

F. EXPERIMENTAL SETUP

Experiments were performed for the three categories of models namely, LSTM-based, Transformer-based, and LLM models. The three LSTM-based models described in the aforementioned sections i.e. AE-LSTM, AT-LSTM, and ATAE-LSTM, were trained on each of the given datasets. In addition, ReduceLROnPlateau[17] callback was used with a patience level of 2 to monitor the validation accuracy and minimize the overfitting of the models. Moreover, a total of 36 analyses were executed to assess the performance of the trained models under in-domain settings. This entailed subjecting the trained models to evaluation on the same dataset domain that was utilized during the training phase. Afterwards, the most suitable model with the highest performance was used to evaluate the performance for cross-domain settings. These entailed subjecting the trained models to evaluation on the different domains that were not utilized during the training phase. These analyses, conducted both within the domain and across domains, aimed to verify the extent to which LSTM-based models could generalize their sentiment analysis capabilities to unseen aspects or domains.

Apart from LSTM-based models discussed in the above sections, the Python 3.9 version was used to make the API calls to the LLMs(Bard, and GPT-3.5-turbo) to retrieve the sentiments for the provided review and aspect categories or aspect terms. Further, Regular expressions were used to extract the sentiments only and discard any unwanted characters from the API responses. Further, a confusion matrix was used to compare the predicted sentiment list with the actual sentiment list. By comparing the predicted sentiment to the actual sentiment for each given term or category, we can identify specific patterns of misclassification. Moreover, the performance of the above-mentioned models was evaluated using the weighted F-measure and overall accuracy. These measures were used to assign equivalent weight to the irregularly distributed Classes in the collected dataset [78].

IV. RESULTS

This section presents a detailed analysis of the results produced by the various models that are discussed in the aforementioned sections. Table 6 summarized the results produced by all the mentioned models. The subsequent sessions present detailed analyses of all the experiments that were performed on the acquired datasets.

A. PERFORMANCE ANALYSIS OF LSTM-BASED MODELS

This section describes the analyses for both, the in-domain and cross-domain settings for LSTM-based models. The overall accuracy and F-measure$_W$ of 33 analyses (11 datasets

[13]https://openai.com/
[14]https://developers.generativeai.google/products/makersuite
[15]https://openai.com/
[16]https://developers.generativeai.google/products/makersuite
[17]https://keras.io/api/callbacks/reduce_lr_on_plateau/

variants and 3 LSTM-based models) are shown in Table 2 for in-domain settings. The values in bold font show the highest values in these tables. From these tables, the performance variation can be seen among the three models. In all 36 analyses, the highest overall accuracy, and F-measure$_W$ were achieved by the ATAE-LSTM model for ATSA and ACSA tasks for most of the datasets and domains. However, there were only two domains of the DOTSA dataset specifically for restaurants, and hotels domains where AT-LSTM showed the highest performance. Hence, the ATAE-LSTM model was further utilized to execute the analyses for cross-domain settings.

For cross-domain settings, the model trained on one dataset's domain was evaluated on the remaining domains and datasets. These analyses were executed to validate the extent to which the ATAE-LSTM model could generalize its capabilities to unseen domains.

The overall accuracy of ATAE-LSTM for ATSA task with cross-domain settings is shown in Table 3. The results revealed that the model trained in one domain was not providing comparable results for other domains except for a few cases. Remarkably, aside from the YASO dataset, models trained on the Books and Clothing datasets did not perform comparably well in other domains. The YASO dataset's relatively close results can be explained by the fact that a sizeable portion of the dataset is classified as ''movie,'' while another sizeable portion is classified as ''others.''. This similarity in labeling suggests that there may be analogous patterns in word embeddings between book and movie reviews.

On the other hand, it is also noteworthy that the model trained on the hotel domain yielded comparable results for restaurant and hotel domains which are related domains with similarities. Similarly, the model trained on the restaurant domain showed comparable performance for the restaurant and hotel datasets. Hence, it was concluded from the results that there is still a need for more labeled datasets that contain diverse domains that can yield comparable results for various domains such as books, clothes, and others. However, labeling such a huge dataset is still an open research question.

The overall accuracy of ATAE-LSTM with cross-domain settings for the ACSA task is shown in Table 4. The results revealed a similar pattern as of ATSA that is mentioned in the above paragraph. The model trained in one domain did not provide comparable results for other domains. However, it is noteworthy that the model trained on the SemEval16/Restaurant domain did not yield comparable results for the MAMS dataset which also contains the restaurant review dataset. The possible reason for this is the list of categories that are mentioned in these two datasets. Furthermore, the MAMS dataset is more challenging as compared to the SemEval16/Restaurant.

On the other hand, when the model was trained on the MAMS dataset, the accuracy was improved MAMS whereas the accuracy was still low for SemEval16/Restaurant. Hence, it was concluded from the overall results that there is no dataset available in the literature for ACSA tasks that can yield comparable accuracies for various domains. There is a need for a large-scale dataset that contains aspect categories for various domains or a model that can perform well for all the domains without requiring a huge dataset for training.

B. PERFORMANCE ANALYSIS OF FINE-TUNED ABSA MODELS

This section discusses the results of ABSA models that have recently been proposed and fine-tuned on ABSA datasets. DeBERTa-v3-base-ABSA-v1.1 fine-tuned with 180k examples for the ABSA dataset including SemEval challenge of datasets form 2014 and 2016 (Laptop14 and Restaurant14, Restaurant-16) [17], [72], MAMS [13], Television [73], and TShirt [74], [75]. Overall, this model is mainly fine-tuned in the Restaurant, Laptop, Tshirt, and Television domains. Similarly, flan-t5-large-ABSA is fine-tuned on a custom dataset prepared by GPT-4 and verified by a human [77]. However, the domains of GPT-4 generated reviews are not explicitly mentioned. To evaluate the generalizability of these fine-tuned ABSA models, 22 analyses (11 dataset variants and 2 models) were executed. The overall accuracy and F1-score of fine-tuned models are shown in Table 5.

The analysis results indicate that DeBERTa-v3-base-ABSA-v1.1 outperformed the flan-t5-large-ABSA both in terms of overall accuracy and f1-score. Notably, DeBERTa was trained on diverse ABSA datasets, including MAMS, Sentihood, and SemEval16. In contrast, flan-t5 was specifically trained and fine-tuned on a custom dataset that was generated from GPT-4. This implies that the performance of flan-t5 is solely reliant on the quality of the dataset, while DeBERTa's substantial success can be attributed to the utilization of standard datasets during training.

DeBERTa has exhibited exceptional results for the ATSA task, particularly excelling in the Hotels and Restaurant domains across various datasets. Notably, it achieved remarkable overall accuracies of 93.1%, 83.7%, 83.8%, and 84.5%, along with corresponding impressive f1-scores of 91.1%, 79.9%, 83.8%, and 82.1% for the DOTSA/Hotel, DOTSARestaurant, MAMS/Restaurant, and SemEval16/Restaurant datasets, respectively.

Despite these remarkable achievements, it is essential to highlight that while DeBERTa excelled in well-established domains used during its training, its performance was comparatively lower for new or unseen domains, such as Books and Clothing from DOTSA. In addition, the results also revealed that DeBERTa's performance exhibited relative limitations for the ACSA task the target is to detect the sentiment of aspect categories that are not necessarily mentioned in the review. This emphasizes the ongoing necessity for a model capable of generalizing its capabilities to unexplored domains without an extensive reliance on labeled datasets, which can be a laborious undertaking. Hence, LLMs can be utilized to overcome such limitations of the ABSA models

TABLE 2. Evaluation of all LSTM-based models(Accuracy and F1-score), The best-performing models are highlighted in bold for each of the datasets.

Dataset	Domain	Task	AE-LSTM		AT-LSTM		ATAE-LSTM	
			Acc	F1	Acc	F1	Acc	F1
DOTSA	Books	ATSA	64.1	57.0	66.6	53.2	**69.6**	**67.3**
	Clothing	ATSA	63.8	54.1	61.7	54.1	62.8	**60.6**
	Hotels	ATSA	88.0	86.3	**90.2**	**88.2**	86.7	86.4
	Restaurant	ATSA	75.4	64.8	**75.4**	64.8	70.1	**69.9**
MAMS	Restaurant	ACSA	59.3	57.3	43.6	26.5	**63.9**	**63.1**
	Restaurant	ATSA	57.8	57.3	45.4	28.4	**63.3**	**62.9**
SemEval2016	Laptops	ACSA	60.9	62.0	62.1	58.6	**68.7**	**68.8**
	Restaurants	ACSA	69.9	67.3	70.1	70.8	**80.3**	**79.0**
	Restaurants	ATSA	74.9	59.8	74.3	63.4	**80.9**	**80.1**
Sentihood	Neighbourhood	ACSA	79.7	79.3	72.5	60.9	**84.0**	**83.6**
YASO	Multi-domain	ATSA	61.1	63.0	77.5	67.7	**86.2**	**85.7**

TABLE 3. Overall accuracy of ATAE-LSTM for ATSA task with cross-domain setting.

Dataset/ Domain	DOTSA /Books	DOTSA/ Clothing	DOTSA/ Hotels	DOTSA/ Restaurant	MAMS/ Restaurant	SemEval16/ Restaurant	YASO/ Multi
DOTSA/ Books	70.92	52.55	58.68	65.68	32.26	56.46	**72.99**
DOTSA/ Clothing	**61.41**	60.2	56.7	58.32	30.76	54.69	**71.49**
DOTSA/ Hotels	64.13	52.55	85.61	**73.13**	30.23	**74.92**	**76.84**
DOTSA/ Restaurant	44.27	42.58	**63.32**	65.39	31.34	**64.85**	**66.84**
MAMS/ Restaurant	33.42	36.73	50.63	45.6	62.64	53.07	48.87
SemEval16/ Restaurant	55.16	53.57	**75.59**	67.44	36.75	81.53	**81.99**
YASO/ Multi	60.86	61.73	**75.77**	61.73	32.41	72	85.2

TABLE 4. Overall accuracy of ATAE-LSTM for ACSA task with cross-domain setting.

Dataset/ Domain	Sentihood/ Neighborhood	SemEval16/ Restaurant	SemEval16/ Laptop	MAMS
Sentihood/ Neighborhood	83.5	69.49	60.92	30.29
SemEval16/ Restaurant	69.44	80.9	62.79	38.4
SemEval16/ Laptop	58.72	67.63	68.16	32.18
MAMS	41.21	44.7	44.44	63.81

TABLE 5. Evaluation of all fine-tuned ABSA models(Accuracy and F1-score), The best-performing models are highlighted in bold for each of the datasets.

Dataset	Domain	Task	Flan-t5		DeBERTa	
			Acc	F1	Acc	F1
DOTSA	Books	ATSA	68.5	69.6	**74.7**	**70.3**
	Clothing	ATSA	74.0	75.5	**78.6**	**70.6**
	Hotels	ATSA	90.2	90.9	**93.1**	**91.1**
	Restaurant	ATSA	79.5	79.7	**83.7**	**79.9**
MAMS	Restaurant	ACSA	49.4	47.5	**77.4**	**77.2**
	Restaurant	ATSA	52.6	51.0	**83.8**	**83.8**
SemEval2016	Laptops	ACSA	68.8	75.9	**81.9**	**79.0**
	Restaurant	ACSA	85.0	87.4	**83.5**	**80.7**
	Restaurant	ATSA	84.5	87.6	**84.6**	**82.1**
Sentihood	Neighbourhood	ACSA	52.3	66.3	**87.6**	**87.4**
YASO	Multi-domain	ATSA	83.5	90.2	**95.3**	**95.3**

C. PERFORMANCE ANALYSIS OF LLMs

This section discusses the results of PaLM and GPT-3.5-Turbo LLMs for ATSA and ACSA tasks. A comprehensive evaluation of the performance of these LLMs was conducted, encompassing 22 analyses involving 11 dataset variants and 2 models. The overall accuracy and f1-score for both ATSA and ACSA are presented in Table6, Table7, Table8, and Table9 respectively. These tables also include a comparative analysis of LLMs with baseline studies, providing a broader context for understanding the models' performance. The subsequent sections discuss the comparative analysis of LLMs, offering insights into their strengths and areas of improvement.

TABLE 6. Overall accuracy comparison of LLMs for ATSA task.

Dataset	Domain	Baseline	ATAE-LSTM	DeBERTa	PaLM	GPT-3.5
DOTSA[57]	Books	31.9[57]	69.6	**74.7**	**79.2**	71.9
	Clothing	43.1[57]	62.8	**78.6**	**77.6**	69.4
	Hotels	37.1[57]	86.7	**93.1**	**93.1**	87.4
	Restaurant	20.8[57]	70.1	**83.7**	**82.8**	79.5
MAMS[13]	Restaurant	84.5[79]	**63.3**	**83.8**	48.8	**61.8**
SemEval16[17]	Restaurant	90.3[80]	80.9	**84.6**	**93.5**	88.8
YASO[56]	Multi	53.7[56]	86.2	**95.3**	**98.1**	91.3

TABLE 7. F1-score Comparison of LLMs for ATSA task.

Dataset	Domain	Baseline	ATAE-LSTM	DeBERTa	PaLM	GPT-3.5
DOTSA	Books	31.9	67.3	70.3	**74.8**	72.5
	Clothing	43.1	60.6	70.6	**69.7**	68.8
	Hotels	37.1	86.4	91.1	**91.8**	88.6
	Restaurant	20.8	69.9	79.9	**81.7**	81.9
MAMS	Restaurant	83.7	62.9	83.8	35.1	**61.2**
SemEval16	Restaurant	85.6	80.1	82.1	**91.8**	90.5
YASO	Multi	85.5	85.7	95.3	**98.2**	94.6

TABLE 8. Overall accuracy comparison of LLMs for ACSA task.

Dataset	Domain	Baseline	ATAE-LSTM	DeBERTa	PaLM	GPT-3.5
Sentihood[18]	Neighbourhood	93.3[61]	84.0	**87.6**	**90.7**	67.8
SemEval16[17]	Laptop	88.4 [81]	68.7	**81.9**	**91.0**	67.8
	Restaurant	84.0[13]	80.3	**83.5**	**92.7**	81.1
MAMS[13]	Restaurant	74.0 [13]	63.9	**77.4**	55.8	**64.5**

TABLE 9. F1-score comparison of LLMs for ACSA task.

Dataset	Domain	Baseline	ATAE-LSTM	DeBERTa	PaLM	GPT-3.5
Sentihood	Neighbourhood	87.9	83.6	87.4	**90.9**	65.4
SemEval16	Laptop	-	68.8	79.0	**88.7**	75.4
	Restaurant	-	79.0	80.7	**90.9**	85.7
MAMS	Restaurant	-	63.1	77.2	48.4	**64.4**

1) COMPARATIVE ANALYSIS FOR ATSA TASK

The overall accuracy, and f1-score of baseline studies and the LLM models for ATSA tasks are summarized in Table 6 and Table 7. Notably, our investigation reveals that PaLM consistently demonstrates remarkable performance across diverse domains, occasionally surpassing DeBERTa—a model acknowledged for its recent advancements and trained on multiple Aspect-Based Sentiment Analysis (ABSA) datasets. This noteworthy trend underscores PaLM's robustness and competitiveness, especially in comparison to the

latest developments in ABSA models. Specifically, our findings indicate that PaLM yields comparable results to DeBERTa for all domains of DOTSA [57] dataset. Particularly PaLM's efficacy is impressive in domains such as Hotels and Restaurants, where its accuracy often outshines that of DeBERTa. In addition, it also beats the baseline accuracy for all the domains of DOTSA with 93.1% and 82.8% accuracy for hotel and restaurant domains respectively. Furthermore, PaLM yielded exceptional results for SemEval16/Restaurant [17] and YASO [56] datasets, and outshined the performance of DeBERTa as well as the baseline accuracy. Particularly, PaLM yielded 93.5% and 98.1% accuracy for SemEval16/Restaurant and YASO datasets respectively.

However, it is essential to acknowledge that, unlike other domains, PaLM faces challenges in the ''MAMS'' dataset, where its performance is notably lower than DeBERTa. On the other hand, GPT-3.5 yielded better accuracy for the MAMS dataset as compared to PaLM. It exhibits the comparable results that were achieved by the ATAE-LSTM model which is specifically trained on the MAMS dataset. It is also noticeable that the accuracy of the ATAE-LSTM model was very low for cross-domain settings as presented in Table 3. Hence, the utilization of PaLM eliminates the training step by yielding similar results that were achieved by training the model which requires a large set of labeled datasets. While PaLM excels in several domains, its limitations in the ''MAMS'' dataset warrant careful consideration and may suggest areas for improvement or adaptation. Nonetheless, for the MAMS dataset, GPT-3.5 yielded comparable results concerning baseline and ATAE-LSTM. More robust LLMs such as GPT-4 can be explored in the future to achieve significant results for challenging datasets such as MAMS. These findings contribute valuable insights for researchers and practitioners navigating the selection of ABSA models, urging a nuanced evaluation that goes beyond overall performance to encompass domain-specific intricacies.

2) COMPARATIVE ANALYSIS FOR ACSA TASK

The overall accuracy and F1-score comparisons between baseline studies and LLM models for ACSA are presented in Table 8 and Table 9. PaLM LLM consistently exhibits commendable performance across various domains and outperforms the DeBERTa in terms of overall accuracy and f-score. This observation underscores PaLM's robustness and competitiveness for ACSA task as compared to the latest advancements in ABSA models. Particularly PaLM's efficacy is impressive in domains such as Hotels, Restaurants, and neighborhood domains, where its accuracy and f-score outshine that of DeBERTa. In addition, it also beats the baseline accuracy for Sentihood [18], and SemEval16 [17] with 90.7%, 91.0%, and 92.7% accuracy for Sentihood/neighborhood, SemEval16/Laptop and SemEval16/Restaurant domains respectively. Furthermore, in terms of f-score PaLM's performance is also outstanding with 90.9%, 88.7%, and 90.9% f-score for Sentihood/neighborhood, SemEval16/Laptop, and SemEval16/Restaurant domains respectively.

V. DISCUSSION

From the outcomes of our experiments, we derive important key findings and suggestions for ABSA tasks across various domains that are discussed in subsequent sections.

A. LSTM-BASED MODELS

The ATAE-LSTM model consistently outperformed others for in-domain settings. However, challenges in cross-domain settings, especially for models trained on specific domains like Books and Clothing, struggled to perform well in unseen domains. Notably, the model trained on the hotel domain demonstrated comparable results for both restaurant and hotel domains, indicating the shared sentiment expressions between closely related domains. These findings emphasize the importance of addressing domain-specific nuances and prompt further exploration into tailored model architectures. There is a need to develop diverse datasets and explore alternative architectures to improve sentiment analysis model robustness.

B. FINE-TUNED ABSA MODELS

DeBERTa-v3-base-ABSA-v1.1, fine-tuned on diverse ABSA datasets including MAMS [13], Sentihood [18], and SemEval16 [17], outperformed flan-t5-large-ABSA. DeBERTa's extensive training on standard datasets led to superior overall accuracy and F1-score. While excelling in established domains like Hotels and Restaurants for ATSA tasks, DeBERTa faced limitations in new or unseen domains, like Books and Clothing from the DOTSA dataset. This emphasizes the need for ABSA models with broad generalization capabilities across diverse domains without an exhaustive dependence on labeled datasets. Furthermore, the results also highlighted DeBERTa's relative shortcomings in the ACSA task, emphasizing the need for models capable of detecting sentiment in aspect categories not explicitly mentioned in the reviews.

C. LLM

In terms of both, ATSA and ACSA tasks, PaLM consistently exhibits commendable performance, outperforming DeBERTa in both overall accuracy and F1-score across various domains. However, challenges arise in the ''MAMS'', where its performance falls notably behind. This underscores the need for continued refinement to ensure the model's adaptability to various and challenging datasets such as MAMS with multiple aspects and multiple sentiments in a single review. Interestingly, GPT-3.5 exhibits promise, particularly in challenging datasets such as MAMS, where its results compete favorably with specialized models like ATAE-LSTM. Notably, these results are encouraging as they were achieved without prior training or labeled datasets.

The possible reasons for LLMs huge success in the tasks of ATSA and ACSA are listed below:

1) Deep understanding of language: Large volumes of text data are used to train LLMs like PaLM. They can record intricate grammatical constructions, semantic relationships, and nuances of language use. This is crucial in ABSA, where identifying aspects, understanding sentiment polarity in context, and distinguishing between objective and subjective language is essential.
2) Transfer Learning: LLMs are often pre-trained on general language tasks, allowing them to transfer that knowledge to specific tasks like ABSA. This provides a strong foundation for ABSA-specific fine-tuning.
3) Contextual Reasoning: LLMs can analyze text considering the surrounding context. In ABSA, this is vital for understanding the sentiment towards an aspect. For example, ''The phone has a great camera'' expresses positive sentiment, while ''The battery life is terrible, but the camera is great'' requires understanding contrasting opinions within the context of the entire sentence.

VI. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this study, we conducted a comparative analysis of ABSA models, comparing the performance of baseline studies LLMs approaches across various datasets and domains. Our findings highlight the consistent and remarkable performance of the PaLM model, which often surpassed DeBERTa, a model recognized for recent advancements in ABSA and trained on multiple ABSA datasets. PaLM demonstrated commendable efficacy across diverse domains. The model outperformed not only DeBERTa but also baseline accuracy. However, it is crucial to acknowledge the challenges faced by PaLM in the ''MAMS'' dataset, where its performance was notably lower than that of DeBERTa. GPT-3.5, on the other hand, yielded better accuracy for the MAMS dataset, comparable to the results achieved by the ATAE-LSTM model specifically trained on MAMS. These findings underscore the complexity of model performance in different domains and highlight the need for nuanced evaluations beyond overall accuracy. The limitations observed in the ''MAMS'' dataset prompt consideration for potential improvements or adaptations, and the exploration of more robust Language Models, such as GPT-4, may offer promising avenues for future research. Our study contributes valuable insights for researchers and practitioners navigating the selection of ABSA models, emphasizing the importance of domain-specific considerations. Leveraging LLMs like PaLM and GPT-3.5 eliminates the requirement for labeled datasets in ABSA tasks, streamlining the training process and mitigating the need for high-computation machines across domains. While PaLM emerges as a robust performer across various domains, the nuanced challenges posed by specific datasets, like ''MAMS,'' underscore the evolving nature of ABSA tasks. Continued research and exploration of LLMs hold the key to addressing these challenges and further enhancing the field of ABSA. Subsequent research endeavors could explore the evaluation of more advanced LLMs, like GPT-4, to further enhance ABSA capabilities, particularly when dealing with challenging datasets.

REFERENCES

[1] S. Kotha and S. Basu, ''Amazon and ebay: Online retailers as market makers,'' in *The Market Makers*. Oxford, U.K: Oxford Univ. Press, 2011, pp. 155–180.

[2] Y. Tim, L. Cui, and Z. Sheng, ''Digital resilience: How rural communities leapfrogged into sustainable development,'' *Inf. Syst. J.*, vol. 31, no. 2, pp. 323–345, Mar. 2021.

[3] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr., *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Hoboken, NJ, USA: Wiley, 2017.

[4] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, ''Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions,'' *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[5] S. M. Mohammad, ''Sentiment analysis: Detecting valence, emotions, and other affectual states from text,'' in *Emotion Measurement*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 201–237.

[6] B. Liu, ''Sentiment analysis and subjectivity,'' in *Handbook of Natural Language Processing*, vol. 2. London, U.K.: Chapman & Hall, 2010, pp. 627–666.

[7] G. Brauwers and F. Frasincar, ''A survey on aspect-based sentiment classification,'' *ACM Comput. Surveys*, vol. 55, no. 4, pp. 1–37, Apr. 2023.

[8] K. W. Trisna and H. J. Jie, ''Deep learning approach for aspect-based sentiment classification: A comparative review,'' *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2014186.

[9] Y. Yu, X. Si, C. Hu, and J. Zhang, ''A review of recurrent neural networks: LSTM cells and network architectures,'' *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[10] L. Floridi and M. Chiriatti, ''GPT-3: Its nature, scope, limits, and consequences,'' *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020.

[11] M. V. Koroteev, ''BERT: A review of applications in natural language processing and understanding,'' 2021, *arXiv:2103.11943*.

[12] P. R. J. Dhanith and K. S. S. Prabha, ''A critical empirical evaluation of deep learning models for solving aspect based sentiment analysis,'' *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13127–13186, Nov. 2023.

[13] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, ''A challenge dataset and effective models for aspect-based sentiment analysis,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 6280–6285.

[14] Z. Nasim and S. Haider, ''ABSA toolkit: An open source tool for aspect based sentiment analysis,'' *Int. J. Artif. Intell. Tools*, vol. 26, no. 6, Dec. 2017, Art. no. 1750023.

[15] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, ''Adaptive recursive neural network for target-dependent Twitter sentiment classification,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 49–54.

[16] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, ''SemEval-2015 task 12: Aspect based sentiment analysis,'' in *Proc. 9th Int. Workshop Semantic Eval.*, 2015, pp. 486–495.

[17] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, and V. Hoste, ''SemEval-2016 task 5: Aspect based sentiment analysis,'' in *Proc. Workshop Semantic Eval. (SemEval)*, 2016, pp. 19–30.

[18] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, ''SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods,'' in *Proc. 26th Int. Conf. Comput. Linguistics, Technical Papers*, Osaka, Japan, Dec. 2016, pp. 1546–1556. [Online]. Available: https://aclanthology.org/C16-1146

[19] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ''A survey of convolutional neural networks: Analysis, applications, and prospects,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[20] T. Lin, Y. Wang, X. Liu, and X. Qiu, ''A survey of transformers,'' *AI Open*, vol. 3, pp. 111–132, Jan. 2022.

[21] H. Sadr, M. M. Pedram, and M. Teshnehlab, ''Convolutional neural network equipped with attention mechanism and transfer learning for enhancing performance of sentiment analysis,'' *J. AI Data Mining*, vol. 9, no. 2, pp. 141–151, 2021.

[22] X. Wang, F. Li, Z. Zhang, G. Xu, J. Zhang, and X. Sun, "A unified position-aware convolutional neural network for aspect based sentiment analysis," *Neurocomputing*, vol. 450, pp. 91–103, Aug. 2021.
[23] B. T. Do, "Aspect-based sentiment analysis using bitmask bidirectional long short term memory networks," in *Proc. 31St Int. Flairs Conf.*, 2018, pp. 1–6.
[24] E. Setiawan, F. Ferry, J. Santoso, S. Sumpeno, K. Fujisawa, and M. Purnomo, "Bidirectional GRU for targeted aspect-based sentiment analysis based on character-enhanced token-embedding and multi-level attention," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 5, pp. 392–407, Oct. 2020.
[25] J. Zeng, X. Ma, and K. Zhou, "Enhancing attention-based LSTM with position context for aspect-level sentiment classification," *IEEE Access*, vol. 7, pp. 20462–20471, 2019.
[26] H. T. Nguyen and M. Le Nguyen, "Effective attention networks for aspect-level sentiment classification," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 25–30.
[27] C. Yang, H. Zhang, B. Jiang, and K. Li, "Aspect-based sentiment analysis with alternating coattention networks," *Inf. Process. Manag.*, vol. 56, no. 3, pp. 463–478, May 2019.
[28] A. Kumar, P. Gupta, R. Balan, L. B. M. Neti, and A. Malapati, "BERT based semi-supervised hybrid approach for aspect and sentiment classification," *Neural Process. Lett.*, vol. 53, no. 6, pp. 4207–4224, Dec. 2021.
[29] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proc. 22nd Nordic Conf. Comput. Linguistics*, Sep. 2019, pp. 187–196.
[30] Y. Peng, T. Xiao, and H. Yuan, "Cooperative gating network based on a single BERT encoder for aspect term sentiment analysis," *Int. J. Speech Technol.*, vol. 52, no. 5, pp. 5867–5879, Mar. 2022.
[31] H. T. Phan, N. T. Nguyen, and D. Hwang, "Aspect-level sentiment analysis using CNN over BERT-GCN," *IEEE Access*, vol. 10, pp. 110402–110409, 2022.
[32] A. Mewada and R. K. Dewang, "SA-ASBA: A hybrid model for aspect-based sentiment analysis using synthetic attention in pre-trained language BERT model with extreme gradient boosting," *J. Supercomput.*, vol. 79, no. 5, pp. 5516–5551, Mar. 2023.
[33] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.
[34] R. G. Cowell, P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Berlin, Germany: Springer, 2007.
[35] Y. Wang, Q. Chen, J. Shen, B. Hou, M. Ahmed, and Z. Li, "Aspect-level sentiment analysis based on gradual machine learning," *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106509.
[36] J. Wang, B. Xu, and Y. Zu, "Deep learning for aspect-based sentiment analysis," in *Proc. Int. Conf. Mach. Learn. Intell. Syst. Eng. (MLISE)*, 2021, pp. 267–271.
[37] H. Xu, L. Shu, P. Yu, and B. Liu, "Understanding pre-trained BERT for aspect-based sentiment analysis," in *Proc. 28th Int. Conf. Comput. Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain: International Committee on Computational Linguistics, Dec. 2020, pp. 244–250. [Online]. Available: https://aclanthology.org/2020.coling-main.21
[38] W. Luo, W. Zhang, and Y. Zhao, "A survey of transformer and GNN for aspect-based sentiment analysis," in *Proc. Int. Conf. Comput. Inf. Sci. Artif. Intell. (CISAI)*, Sep. 2021, pp. 353–357.
[39] V. Jagtap and K. Pawar, "Analysis of different approaches to sentence-level sentiment classification," *Int. J. Sci. Eng. Technol.*, vol. 2, no. 3, pp. 164–170, 2013.
[40] J. Zhou, J. X. Huang, Q. Chen, Q. V. Hu, T. Wang, and L. He, "Deep learning for aspect-level sentiment classification: Survey, vision, and challenges," *IEEE Access*, vol. 7, pp. 78454–78483, 2019.
[41] B. Mahesh, "Machine learning algorithms—A review," *Int. J. Sci. Res.*, vol. 9, pp. 381–386, Jan. 2020.
[42] G. I. Webb, E. Keogh, and R. Miikkulainen, "Naïve Bayes," *Encyclopedia Mach. Learn.*, vol. 15, no. 1, pp. 713–714, 2010.
[43] W. S. Noble, "What is a support vector machine?" *Nature Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006.
[44] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. Pharmaceutical Biomed. Anal.*, vol. 22, no. 5, pp. 717–727, Jun. 2000.
[45] R. Varghese and M. Jayasree, "Aspect based sentiment analysis using support vector machine classifier," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2013, pp. 1581–1586.
[46] Y. Xu, K. Hong, J. Tsujii, and E. I.-C. Chang, "Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 824–832, Sep. 2012.
[47] Y. Zhang, J. Wang, and X. Zhang, "Conciseness is better: Recurrent attention LSTM model for document-level sentiment analysis," *Neurocomputing*, vol. 462, pp. 101–112, Oct. 2021.
[48] M. R. Chavez, T. S. Butler, P. Rekawek, H. Heo, and W. L. Kinzler, "Chat generative pre-trained transformer: Why we should embrace this technology," *Amer. J. Obstetrics Gynecol.*, vol. 228, no. 6, pp. 706–711, Jun. 2023.
[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[50] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," 2019, *arXiv:1905.05950*.
[51] R. Dale, "GPT-3: What's it good for?" *Natural Lang. Eng.*, vol. 27, no. 1, pp. 113–118, 2021.
[52] V. Cohen and A. Gokaslan, "OpenGPT-2: Open language models and implications of generated text," *XRDS, Crossroads, ACM Mag. Students*, vol. 27, no. 1, pp. 26–30, Sep. 2020.
[53] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?" *Library Hi Tech. News*, vol. 40, no. 3, pp. 26–29, May 2023.
[54] S. Chumakov, A. Kovantsev, and A. Surikov, "Generative approach to aspect based sentiment analysis with GPT language models," *Proc. Comput. Sci.*, vol. 229, pp. 284–293, Jan. 2023.
[55] D. Magdaleno, M. Montes, B. Estrada, and A. Ochoa-Zezzatti, "A GPT-based approach for sentiment analysis and bakery rating prediction," in *Proc. Mexican Int. Conf. Artif. Intell.* Mérida, México: Springer, 2023, pp. 61–76.
[56] O. Toledo-Ronen, M. Orbach, Y. Katz, and N. Slonim, "Multi-domain targeted sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2022, pp. 2751–2762.
[57] Y. Luo, H. Cai, L. Yang, Y. Qin, R. Xia, and Y. Zhang, "Challenges for open-domain targeted sentiment analysis," 2022, *arXiv:2204.06893*.
[58] H. Luo, T. Li, B. Liu, and J. Zhang, "DOER: Dual cross-shared RNN for aspect term-polarity co-extraction," 2019, *arXiv:1906.01794*.
[59] J. Machacek, "BUTknot at SemEval-2016 Task 5: Supervised machine learning with term substitution approach in aspect category detection," in *Proc. 10th Int. Workshop Semantic Eval.*, 2016, pp. 301–305.
[60] S. U. S. Chebolu, P. Rosso, S. Kar, and T. Solorio, "Survey on aspect category detection," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–37, Jul. 2023.
[61] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Minneapolis, MN, USA, 2019, pp. 380–385. [Online]. Available: https://aclanthology.org/N19-1035
[62] M. S. Akhtar, T. Garg, and A. Ekbal, "Multi-task learning for aspect term extraction and aspect sentiment classification," *Neurocomputing*, vol. 398, pp. 247–256, Jul. 2020.
[63] M. Orbach, O. Toledo-Ronen, A. Spector, R. Aharonov, Y. Katz, and N. Slonim, "YASO: A targeted sentiment analysis evaluation dataset for open-domain reviews," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9154–9173.
[64] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The multilingual Amazon reviews corpus," 2020, *arXiv:2010.02573*.
[65] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, vol. 1631, Jan. 2013, pp. 1631–1642.
[66] K. Ganesan, C. Zhai, and J. Han, "Opinosis: A graph based approach to abstractive summarization of highly redundant opinions," in *Proc. 23rd Int. Conf. Comput. Linguist. (Coling)*, Aug. 2010, pp. 340–348.
[67] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[68] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.

[69] L. Bao, P. Lambert, and T. Badia, "Attention and lexicon regularized LSTM for aspect-based sentiment analysis," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 253–259.
[70] H. Yang, C. Zhang, and K. Li, "PyABSA: A modularized framework for reproducible aspect-based sentiment analysis," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2023, pp. 5117–5122.
[71] H. Yang, B. Zeng, M. Xu, and T. Wang, "Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning," 2021, *arXiv:2110.08604*.
[72] D. Kirange, R. R. Deshmukh, and M. Kirange, "Aspect based sentiment analysis SemEval-2014 task 4," *Asian J. Comput. Sci. Inf. Technol.*, vol. 4, Aug. 2014.
[73] T. Cooray, G. Perera, A. Kugathasan, and J. Alosius, "Aspect-based sentiment analysis: Movie and television series reviews," *Proc. SPIE*, vol. 11766, pp. 615–620, Mar. 2021.
[74] S. Rajapaksha and S. Ranathunga, "Aspect detection in sportswear apparel reviews for opinion mining," in *Proc. Moratuwa Eng. Res. Conf. (MERCon)*, Jul. 2022, pp. 1–6.
[75] R. Mukherjee, S. Shetty, S. Chattopadhyay, S. Maji, S. Datta, and P. Goyal, "Reproducibility, replicability and beyond: Assessing production readiness of aspect based sentiment analysis in the wild," in *Advances in Information Retrieval*. Springer, 2021, pp. 92–106.
[76] B. Zeng, H. Yang, R. Xu, W. Zhou, and X. Han, "LCF: A local context focus mechanism for aspect-based sentiment classification," *Appl. Sci.*, vol. 9, no. 16, p. 3389, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/16/3389
[77] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, and S. Brahma, "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.
[78] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.
[79] X. Bai, P. Liu, and Y. Zhang, "Investigating typed syntactic dependencies for targeted sentiment classification using graph attention neural network," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 503–514, 2021.
[80] H. Yang and K. Li, "Improving implicit sentiment learning via local sentiment aggregation," 2021, *arXiv:2110.08604*.
[81] K. Scaria, H. Gupta, S. Goyal, S. A. Sawant, S. Mishra, and C. Baral, "InstructABSA: Instruction learning for aspect based sentiment analysis," 2023, *arXiv:2302.08624*.

NIMRA MUGHAL received the master's degree (Hons.) in computer science from Sukkur IBA University, Sukkur, Pakistan, in 2022. Currently, she holds the position of a Researcher in artificial intelligence with Sukkur IBA University. In addition to her role in artificial intelligence, she is a Visiting Faculty Member with the Computer Science Department, Sukkur IBA University. Prior to joining Sukkur IBA University, she was with reputed colleges as a Subject Specialist in computer science. She has sound experience in teaching and research. Her research interests include deep learning, machine learning, text mining, natural language processing, and computer vision.

GHULAM MUJTABA received the master's degree (Hons.) in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has been an Associate Professor with Sukkur IBA University, Sukkur, Pakistan, since 2006. Prior to joining Sukkur IBA University, he was with a well-known software house in Karachi for four years. He has vast experience in teaching and research. He has published several articles in academic journals indexed in well-reputed databases, such as ISI and Scopus. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval. He has received the Gold Medal for the master's degree.

SARANG SHAIKH received the master's degree in computer science from Sukkur IBA University, Pakistan, in 2020. He is currently pursuing the Ph.D. degree with the Department of Information Security and Communication Technology, Norwegian University of Science and Technology (NTNU), Norway. Prior to joining NTNU University, he was a Visiting Lecturer with Sukkur IBA University, Pakistan. He is the author of several papers published in international journals. His research interests include applied research in the field of artificial intelligence, NLP, machine learning, deep learning, and learning technologies. He has served as a Reviewer for IEEE Access.

AVEENASH KUMAR received the bachelor's degree from Sukkur IBA University, Pakistan, in 2023. He is currently a Data Scientist with Learners.ai. His research interests include deep learning, natural language processing, and signal processing.

SHER MUHAMMAD DAUDPOTA received the master's and Ph.D. degrees from the Asian Institute of Technology, Thailand, in 2008 and 2012, respectively. He is currently a Professor of computer science with Sukkur IBA University, Pakistan. Alongside his computer science contribution, he is also a Quality Assurance Expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of the Higher Education Commission in the role of an Educational Quality Reviewer. He is the author of more than 35 peer-reviewed journals and conference publications. His research interests include deep learning, natural language processing, video, and signal processing.